LIP FEATURES AUTOMATIC EXTRACTION

M. Liévin and F. Luthon

Signal and Image Laboratory, Grenoble National Polytechnical Institute,
LIS, INPG, 46 av. Félix-Viallet, 38031 Grenoble Cedex, France
email : {lievin-luthon}@tirf.inpg.fr fax : +33 (0)4 76 57 47 90

ABSTRACT

An algorithm for speaker's lip segmentation and features extraction is presented in this paper. A color video sequence of speaker's face is acquired, under natural lighting conditions and without any particular make-up. First, a logarithmic color transform is performed from RGB to HI (hue, intensity) color space. Second, a statistical approach using Markov random field modeling determines red hue prevailing region and motion in a spatiotemporal neighborhood. Third, the final label field is used to extract ROI (Region Of Interest) and geometrical features.

1. INTRODUCTION

It is commonly observed that visual information provides a precious help to the listener under degraded acoustical conditions [1]. Visual cues are effectively used by human beings to improve speech intelligibility. The motivation of the present work is to extract information for automatic audiovisual speech recognition (ASR), videoconferencing and speaker's face synthesis.

Many approaches have been proposed in this area, some are based on gray-level analysis, others on color analysis. Several groups are working with template models based on dynamic contours (*e.g.* Dalton in [6]), active shape models (*e.g.* Luettin in [6]), deformable templates (*e.g.* Silsbee in [6])... Some of them impose strong constraints, like blue make-up [1] or adapted illumination conditions [7]. The previous work [5], which used a linear color transform under regular illumination, needed a large neighborhood and complex label fields to segment the lips.

Here, an algorithm is proposed for lip motion detection under natural conditions, the only requirement being that a micro-camera is mounted on a light helmet worn by the speaker so that it is fixed w.r.t. the head (Fig. 1). The video sequence contains the region of the face spanning from chin to nostrils. The purpose of the process is to obtain, in the mouth region, lip motion and red hue label fields. Lip geometrical features are finally extracted from these label fields (Fig. 1). The processing of each image is divided into three stages:

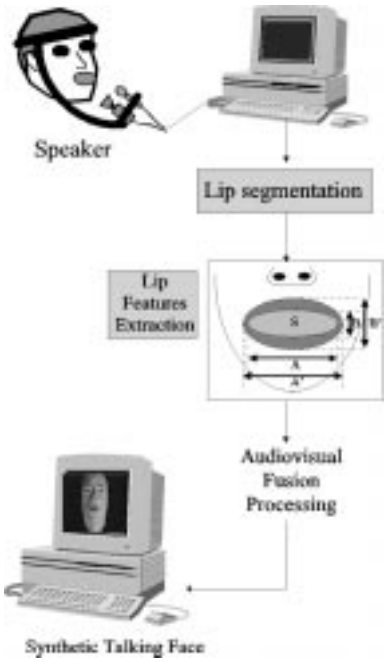

Figure 1: Context of lipreading: from sequence of speaker images, geometrical lip features provide parameters for talking face synthesis.

- Logarithmic color-space transform, *RGB* to *HI*.
- Lip segmentation by an algorithm based on an MRF modeling framework.
- ROI (Region Of Interest) detection and features extraction.

2. LOGARITHMIC COLOR TRANSFORM

First, an *RGB* image sequence (8 bits/pixel/color) of mouth movements is acquired at standard video-rate (25 images/s in Europe), with no particular make-up or lighting. Color-based approaches often use an *RGB* to *HSI* (hue, saturation, intensity) transform to gain independance from illumination brightness [2]. This kind of transform exhibits poor results in noisy conditions (*e.g.* acquisition with a micro-camera video system). Therefore, a logarithmic image processing model [4] is applied here to enhance hue computation. This model satisfies the saturation characteristics of the human visual system and the algebraic operations are justified from a physical point of view.

The logarithmic image processing model is a mathematical framework based on the gray tone function i representation of the intensity I at a pixel $s = (x, y)$. $I(s) \rightarrow i(s) = 256(1 - \frac{I(s)}{I_0})$ where I_0 is a constant representing the incident intensity.

To obtain robust hue observation, the logarithmic difference between *red* and *green* components is computed (specific algebraic operations are defined in [4]) . This operation provides lip detection robust to lighting conditions and variable illumination. Indeed, red color usually prevails in lip areas and green values are close to the luminance component. So, their difference is adapted to detect lip regions. Moreover, considering I_0 close to the maximum value of white (256), the logarithmic difference becomes a ratio between R and G components (Eq. 1) (Fig. 1.

$$H = 256 \times \frac{G}{R} \quad \text{and} \quad I = \frac{R+G+B}{3} \tag{1}$$

Figure 2: *Top*: 5 typical images of luminance sequence; *Bottom*: the corresponding hue sequence.

3. THE LIP-MAD (MOTION AUTOMATIC DETECTION) ALGORITHM

3.1. Observations and Labels

To detect lip regions, motion information is combined with red hue. From the HI color space, two kinds of observations o are derived: a temporal observation $fd(s)$ which is a simple difference between the luminance of two consecutive images, and a spatial observation $h(s)$ computed from the hue.

$$fd(s) = |I_t(s) - I_{t-1}(s)| \tag{2}$$

where $I(s)$ represents the intensity (or luminance) at pixel s.

$h(s)$ consists in filtering the hue value $H(s)$ at pixel s with a parabola centered on the mean value of lip hue H_m, determined beforehand on the first image. The standard deviation of the hue value Δ_H is currently an heuristic parameter (typ. $\Delta_H = 8$).

$$h(s) = \left[256 - \left(\frac{H(s) - H_m}{\Delta_H}\right)^2\right] \times 1_{|H(s)-H_m| \leq 16.\sigma} \tag{3}$$

the notation $1_{condition}$ denotes a binary function which takes the value 1 is the condition is true, 0 otherwise.

An intial label field L_t^0 combining four labels is then defined, these labels corresponding to four typical configurations of motion and hue values (Table 1). The two observations are initially thresholded (thresholds θ_h and θ_{fd}). The thresholds are heuristic but independent of the sequence (typ. $\theta_{fd} = 10$ and $\theta_h = 200$) (Fig. 3).

Initial configuration		motion	red hue	
$fd(s)$	$h(s)$	$m(s)$	$r(s)$	l_s
$< \theta_{fd}$	$< \theta_h$	0	0	b_0
$> \theta_{fd}$		1		b_1
$< \theta_{fd}$	$> \theta_h$	0	1	a_0
$> \theta_{fd}$		1		a_1

Table 1: Low-level information $m()$ and $r()$ and the four corresponding initial labels l_s.

3.2. Neighborhood and MRF Framework

This label field is supposed to follow the main MRF (Markov Random Field) property related to the *spatiotemporal neighborhood* structure η, given in Fig. 4, *i.e.* the label l_s of a pixel s depends only on the labels of its neighbors $n \in \eta(s)$.

Assuming the equivalence between MRFs and Gibbs distributions, Maximizing the A Posteriori probability (MAP criterion) of the label field is equivalent to minimizing a global energy function [3] :

$W(S) = \sum_{o \in \{fd,h\}} U_o(S) + U_m(S)$ where U_o and U_m represent respectively the *attachment energies* (expressing the link between labels and observations, Eq.

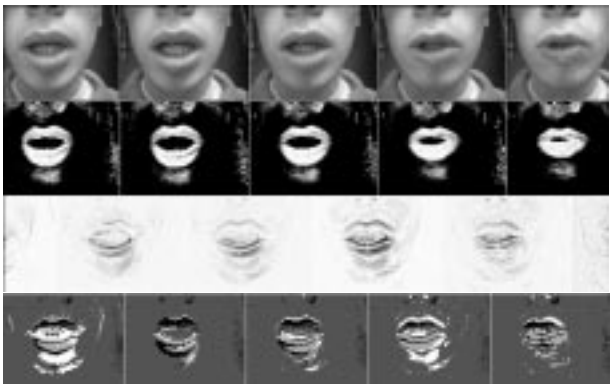

Figure 3: *From top to bottom*: sequence of luminance images; sequence of red prevailing observation; sequence of temporal observation; sequence of initial labels. *The 4 labels are shown in grey levels (from white to black: b_1, a_1, b_0, a_0).*

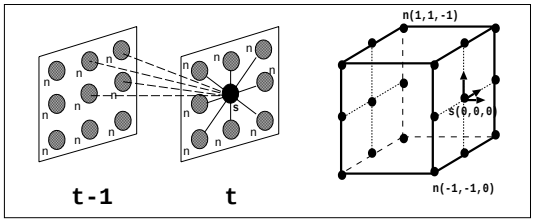

Figure 4: *Left*: Spatiotemporal neighborhood structure η with binary cliques $c = (s, n)$. s is the current pixel (in black), n is any spatiotemporal neighbor of s (in gray);*Right*: corresponding elementary cube C_{xyt}

4) and the *model energy* (associated to spatial and temporal a priori modeling) (Eq. 5) over the image S.

$$U_o(S) = \sum_{s \in S} \left[\frac{[o_s - \psi_o(l_s)]^2}{2\sigma_o^2} \right] \tag{4}$$

where ψ_o is an attachment function, defined in Table 2 and σ_o^2 is the corresponding variance which is estimated on line.

l_s	a_0	b_0	a_1	b_1
ψ_h	$\Psi_h(s)$	0	$\Psi_h(s)$	0
ψ_{fd}	0	0	$\Psi_{fd}(s)$	$\Psi_{fd}(s)$

Table 2: Computation of ψ_o.
($\Psi_o(s) = \frac{1}{card(S_o)} \sum_{s \in S_o} o_s$ where $S_o = \{s \in S / o_s > \theta_o\}$)

The *a priori* model energy is defined as a sum of interaction potential functions over the neighborhood:

$$U_m(S) = \sum_{s \in S} \left[\sum_{n \in \eta(s)} V_{st}(l_n, l_s) \right] \tag{5}$$

The spatiotemporal potential function V_{st} is defined as the inverse of the Euclidian distance between two neighbors in C_{xyt}. The distance integrates two elementary potentials β_s and β_t as compression scales (Eq. 7).

$$V_{st} = \frac{1}{\sqrt{\left(\frac{\delta_x}{\beta_x}\right)^2 + \left(\frac{\delta_y}{\beta_y}\right)^2 + \left(\frac{\delta_t}{\beta_t}\right)^2}} \tag{6}$$

$$= \frac{\beta_s \beta_t}{\sqrt{\beta_t^2 \left(\delta_x^2 + 4\delta_y^2\right) + \beta_s^2 \delta_t^2}} \tag{7}$$

where $\overrightarrow{(s,n)} = (\delta_x, \delta_y, \delta_t)$ and $\delta \in \{-1; 0; 1\}$

We use $\beta_x = 2.\beta_y = \beta_s$ to put emphasize on horizontal configurations. The elementary potentials β_s and β_t are defined to constrain the model respectively to spatial homogeneity of labels and temporal homogeneity of hue when no motion is detected (Eq. 8 and Table 3).

$$\beta_s(l_s, l_n) = 1 + |r(s) - r(n)| + |m(s) - m(n)| \tag{8}$$

$r(n)$	0	1	0	1	1	0	0	1
$r(s)$	0	1	1	0	0	1	0	1
$m(s)$	0	0	1	1	0	0	1	1
β_t	1	1	1	1	2	2	2	2

Table 3: Computation of elementary temporal potential β_t

3.3. The relaxation algorithm

An iterative deterministic algorithm (ICM : Iterated Conditional Modes) is implemented to compute the minimum energy at each site (Eq. 9), starting from the initial label configuration L_t^0. Fig. 5 shows the complete processing scheme.

$$W(S) = (\lambda . U_h(S) + U_{fd}(S)) + \alpha . U_m(S) \tag{9}$$

where λ is a weighting coefficient between the two attachment energies (typ. $\lambda = 1$) and α is a weighting coefficient between the attachment energies and the model energy (typ. $\alpha = 20$).

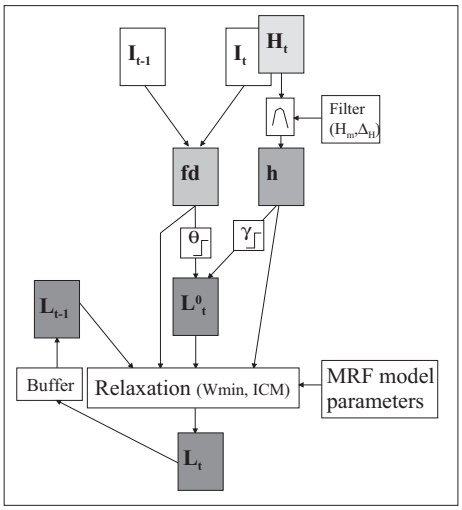

Figure 5: Lip-MAD block diagram

3.4. Results

After a few iterations (typ. 5) on the field, convergence is achieved. One obtains homogeneous red hue and lip motion fields. Thanks to the motion information fd, the lip border is accurately detected and corresponds to the lip shape (Fig. 6). The algorithm tends to keep lip motion labels only if they contribute to the segmentation.

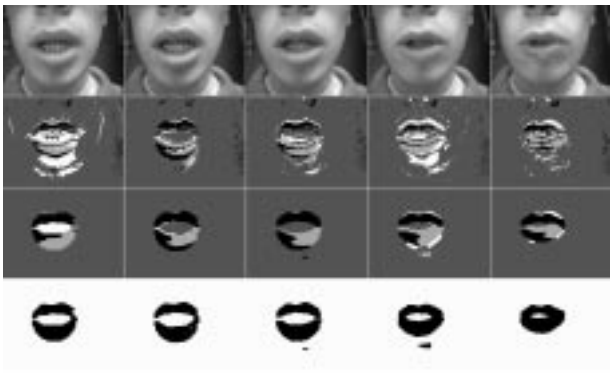

Figure 6: *From top to bottom*: sequence of luminance images; initial labels; label fields after relaxation. *The 4 labels are shown in grey levels (from white to black:* b_1, a_1, b_0, a_0*)*; sequence of hue relevant label images (a_0 and a_1)

4. LIP FEATURES EXTRACTION

4.1. Lip red hue labels and ROI

From the final label fields, one can extract lip red hue relevant label (a_0 and a_1) (Fig. 7). Those results are shown with a ROI evaluated on line.

Figure 7: *Top*: sequence of luminance images ;*Bottom*: sequence of final lip hue fields with ROI superposed on the luminance.

The robustness and the good quality of the label fields allows the lip-MAD algorithm to extract dynamically the ROI and geometrical lip features (Fig. 8). x_0 is the horizontal coordinate of the mean weight of

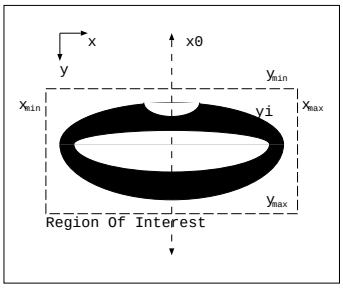

Figure 8: Geometrical lip model

the red hue labels. The ROI (*e.g.* mouth location) is evaluated by maximizing a cost function , (S) (Eq. 10). The ROI needs 4 coordinates to be defined:

$ROI = \{x_{min}, y_{min}, x_{max}, y_{max}\}$.

This values are computed to find the maximum of , (S).

$$, (S) = \frac{\left(\sum_{s \in ROI} 1_{r(s)=1}\right)^2}{surface(ROI)} \qquad (10)$$

where $surface(ROI) = (x_{max} - x_{min})(y_{max} - y_{min})$. The notation $1_{condition}$ denotes a binary funtion (*see* Eq. 3).

Different sequences have been tested, some with a soft natural red make-up (Fig. 9), others with very poor lighting conditions without any make-up (Fig. 10). It shows the robustness of the algorithm to the variability of the lighting conditions.

Figure 9: *Top*: sequence of luminance images with soft red make-up;*Bottom* sequence of final red hue fields with ROI superposed on the luminance.

Figure 10: *Top*: sequence of luminance images with no lighting supply;*Bottom* sequence of final red hue fields with ROI superposed on the luminance.

4.2. Geometrical lip features

Polynomial equations are used for modeling both the internal and external contours. The lip are quite symmetrical. So, the equation of a border i is:

$y_i(x) = y_i(x_0) + \alpha_i(x - x_0)^2$ where x_0 is considered as the mouth's axis. Five contours are sufficient to model the lip borders. First, the internal border, and next the external border, are tracked on the hue relevant labels. The coefficients $y_i(x_0)$ and α_i are computed, by MSE (Mean Squared Error), from the coordinates of the sites detected. One obtains finally the shape of the lips (Fig. 11). This results can be considered as an intermediate processing stage for a further lip parameters fusion with audio information.

Figure 11: *Top*: sequence of luminance images;*Bottom* Lip shape extraction

5. CONCLUSION

The LIP-MAD algorithm emphasizes the need of a good transformation from *RGB* to another color-space. We choose a logarithmic transformation close to the characteristics of the human visual system in order to enhance the subsequent segmentation. Secondly, the algorithm integrates temporal with spatial information. Processing together several images consecutive in time is a way to improve the quality of the contours, often elusive on speaker's lips. Finally, the good quality of the final fields allows an automatic *on line* ROI extraction and geometrical lip features estimation.

We need to process more sequences to develop an automatic mean hue lip H_m estimation (presently, the only speaker dependant parameter). Some more difficult cases need to be studied, like faces with beard or colored people faces.

The proposed algorithm requires less than 10 iterations until convergence (about 2 sec. on SunUltra1). Therefore, the implementation of the algorithm on a programmable video processor is now under study, to achieve processing at video-rate (25 images/sec.).

6. REFERENCES

[1] C. Benoît, M.T. Lallouache, and T. Mohamadi. A set of french visemes for visual speech synthesis. In *Talking Machines: Theories, Models and Designs*, pages 485–504. Elseviers Science Publishers, 1992.

[2] T. Coianiz and L. Torresani. Analysis and encoding of lip movements. In *Audio and Video-based Person Authentification*, pages 53–60, Crans-Montana, Switzerland, March 1997. First International Conference AVBPA'97.

[3] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Analysis Machine Intell.*, 6(6):721–741, November 1984.

[4] M. Jourlin and J-C. Pinoli. Image dynamic range enhancement and stabilization in the context of the logarithmic image processing model. *Traitement du Signal*, 41(2):225–237, January 1995. in French.

[5] F. Luthon and M. Liévin. Lip motion automatic detection. In *Scandinavian Conference on Image Analysis*, volume 1, pages 253–260, Lappenranta, Finland, June 1997.

[6] D. Stork and M. Hennecke. *Speechreading by Humans and Machines*, volume 150. Springer-Verlag, Berlin, 1996.

[7] M. Vogt. Interpreted multi-state lip models for audio-video speech recognition. In *Proceedings of the Audio-Visual Speech Processing, Cognitive and Computational Approaches Workshop*, ISSN 1018-4554, Rhodes (Greece), September 1997.