UNSUPERVISED LIP SEGMENTATION UNDER NATURAL CONDITIONS

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 unsupervised algorithm for speaker's lip segmentation 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 and sequence dependant parameters are evaluated. Second, a statistical approach using Markov random field modeling segment mouth shape using red hue predominant region and motion in a spatiotemporal neighborhood. Simultaneously, a Region Of Interest (ROI) is automatically extracted. Third, the speaker's lip shape is extracted from the final hue field with good quality results in this challenging situation.

1. INTRODUCTION

It is commonly observed that visual information provides a precious help to the listener under degraded acoustical conditions [1]. The motivation of the present work is to extract visual information for automatic speech recognition (ASR), videoconferencing and speaker's face synthesis under natural lighting conditions with few assumptions.

Some approaches proposed in this area are based on gray level analysis (*e.g.* Luettin in [5]). Others use color analysis but need to determine optimal values of some parameters (*e.g.* Coianiz in [5]). Strong assumptions are required on the skin hue parameters and the mouth location [6], therefore the skin hue region is often determined manually beforehand.

The previous work [4] used a segmentation to locate the mouth before estimating lip geometrical features, some of the segmentation parameters were determined beforehand. Here, an algorithm is proposed for unsupervised lip shape extraction and mouth location under natural conditions, the 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. The RGB video sequence (8 bits/color/pixel) contains the region of the face spanning from chin to nostrils. The purpose of the process is to obtain the mouth shape using red hue label fields and motion information. The processing is divided into three stages:

1: Logarithmic color-space transform, *RGB* to *HI*.
2: Estimation of sequence dependant parameters:
Computation of the mean value of the lip hue H_{lip}.
Estimation of the noise on the motion information.
3: Spatiotemporal segmentation of the lip and ROI estimation.

2. PARAMETER ESTIMATION

2.1. Logarithmic color transform

Color–based approaches often use color angles methods (HSI) for illuminant–invariant recognition. Color shifts can be well categorized with angles if camera sensors are sufficiently narrowband. But, in our application, we deal with a mono-CCD camera which gives poor results with angular transforms (noisy conditions). Moreover, G and B channels seem to be correlated in the red region (where R is preponderant). From the RGB color space, we use only two dimensions R and G under the assumption that red prevails in face areas and specially in lip areas. We define the chromacity histogram (R,G) as the non-normalized projection of the RGB color space. The typical histogram of a face sample is shown Fig. 1. Two regions with a specific angular direction appear.

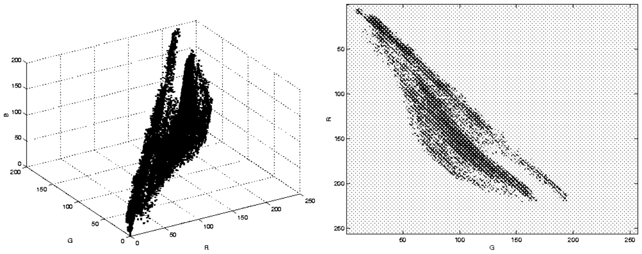

Figure 1: *Left*: 3D RGB histogram of a face under natural illumination; *Right*: 2D RG chromacity histogram.

To obtain a robust hue observation to the lighting conditions, we compute the hue in a mathematical framework based on a logarithmic image processing model [3]. The intensity I of an image is represented by its associated gray tone function $\boldsymbol{i}$ (Eq. 1). This model satisfies the saturation characteristics of the human visual system and is justified from a physical point of view. Specific algebraic and functional operations are redefined in a vectorial structure, like $\oplus$ and $\ominus$ as respectively the addition and the opposite of a gray tone function. The difference between $\boldsymbol{g}$ and $\boldsymbol{f}$, respectively logarithmic tone of the intensity G and F, is given Eq. 2.

$$\begin{aligned} \boldsymbol{i} &= M\left(1-\frac{I}{I_0}\right) & (1)\\ \boldsymbol{f} \oplus (\ominus \boldsymbol{g}) &= M\frac{\boldsymbol{f}-\boldsymbol{g}}{M-\boldsymbol{g}} & (2)\\ \boldsymbol{h} = \boldsymbol{g} \oplus (\ominus \boldsymbol{r}) &= M\left(1-\frac{G}{R}\right) & (3) \end{aligned}$$

Considering the illuminance I_0 close to the maximum value of white M, the logarithmic transform becomes $\boldsymbol{i} = M - I$. We define $\boldsymbol{h}$ as the logarithmic hue tone of H, difference of $\boldsymbol{g}$ and $\boldsymbol{r}$, logarithmic color tone of G and R (Eq. 3). The logarithmic difference becomes a ratio between R and G components $H = M \times \frac{G}{R}$. Finally, from the RGB color space, a HI logarithmic color space (Fig. 2) is defined considering $M = 256$ and the intensity I as the mean value of the R,G and B components (Eq. 4).

$$H = 256 \times \frac{G}{R} \quad \text{and} \quad I = \frac{R+G+B}{3} \qquad (4)$$

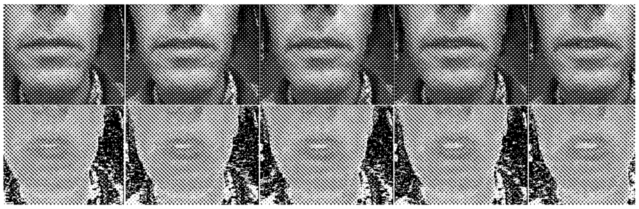

Figure 2: *Top*: 5 typical images of luminance sequence; *Bottom*: the corresponding hue sequence.

2.2. Observations

To detect lip regions, motion information is combined with red hue. From the HI color space, two kinds of observations o are derived, defined to be in the same range $[0 \cdots 255]$ as the image quantification (8 bits). First, a hue observation $h(s)$ consists in filtering the hue value $H(s)$ at pixel s with a parabola centered on the mean value of lip hue H_{lip} with a standard deviation of the hue value Δ_H (Eq. 5).

$$h(s) = \left[256 - \left(\frac{H(s) - H_{lip}}{\Delta_H}\right)^2\right] \times 1_{\frac{|H(s)-H_{lip}|}{\Delta_H} \leq 16} \qquad (5)$$

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

Second, a temporal observation $fd(s)$ is defined as the unsigned difference between the luminance of two consecutive images (Eq. 6). $I(s)$ represents the intensity (or luminance) at pixel s.

$$fd(s) = |I_t(s) - I_{t-1}(s)| \qquad (6)$$

2.3. Hue and motion estimation

The hue segmentation needs three estimated parameters to be unsupervised : H_{lip}, Δ_H, θ_h. Due to the chosen expression of $h(s)$, the tresholded hue field is defined by Eq. 7, expressing the link between Δ_H and θ_h.

$$h(s) > \theta_h \Leftrightarrow |H(s) - H_{lip}| < \Delta_H \sqrt{256 - \theta_h} \qquad (7)$$

The hue histogram $\{p_{i,h}, i \in [0 \cdots 255]\}$ is an useful representation of the hue distribution over the image. We can detect two modes, one for the face, the second for the lip. But, in natural conditions (no make-up), the two modes are mixed (Fig. 3). In order to estimate H_{lip} accurately, a specific hue $H_{\delta_h(face,lip)}$ is defined in Eq. 8. H_{lip} is defined by the Eq. 9.

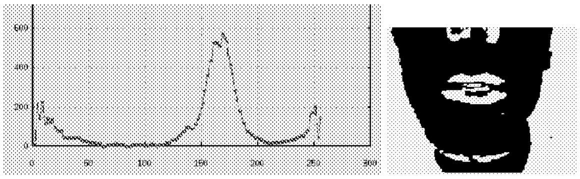

Figure 3: *Left*: histogram of hue image;*Right*: unsupervised segmentation of hue face.

$$\begin{aligned} H_{\delta_h}(S) &= \frac{card(S)}{\delta_h} \sum_{i \in \delta_h} (p_{i,h}(S) \times i) & (8)\\ H_{lip} &= H_{\delta_h(lip)}(S_s) & (9) \end{aligned}$$

δ_h corresponds to the appropriate interval and S_s represents the image S after face segmentation. The processing respects the following steps:

1: Estimate H_{face} using $\{p_{i,h}, i \in [0 \cdots 255]\}$ computed over the hue image.
2: Segment the first hue image with a basic spatial MRF segmentation via the hue transform observation
3: Evaluate H_{lip} using $\{p_{i,h}, i \in [0 \cdots 255]\}$ computed over the segmented hue image.

$\delta_h(lip)$ and $\delta_h(face)$, camera dependant parameters, are independant from the speaker and the lighting conditions. They can be estimated by camera calibration. Currently, the range of δ_h is the result of the statistical distribution of manually estimate over caracteristic natural conditions. The selected range for $\delta_h(face, lip)$ is $[100 \cdots 200]$ and for $\delta_h(lip)$, $[100 \cdots 150]$. This range corresponds to the red predominant region. The equation 7 is then respected with $\theta_h = 192$ and $\Delta_H = 6$.

The algorithm requires an appropriate threshold θ_{fd} to suppress the camera noise without cutting significant temporal changes. In the previous work [4], this threshold was determined before segmentation by hand. We compute here the entropy $E_{fd}(S)$ over an image difference (Eq. 11). This gives the level of noise from which we can deduce the value of θ_{fd} (Eq. 12). The thresholded motion field is then defined by $fd > \theta_{fd}$.

$$p_{i,o}(S) = \frac{1}{card(S)} \sum_{s\in S} 1_{(o(s)=i)} \quad (10)$$

$$E_o(S) = - \sum_{i\in[0\cdots 255]} p_{i,o}(S) log_2(p_{i,o}(S)) \quad (11)$$

$$\theta_{fd}(S) = 2^{E_{fd}(S)} \quad (12)$$

where $p_{i,o}(S)$ represents the probability of level i in the observation o over image S.

The thresholded fields appear non homogeneous and noisy (Fig. 4). Therefore, we need a statistical relaxation to segment more accurately the lip.

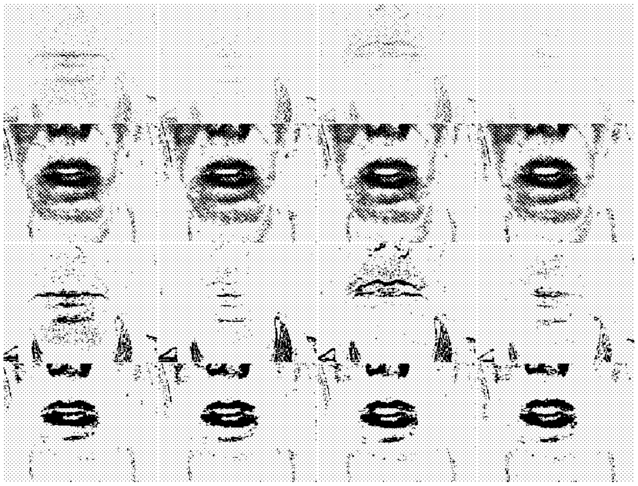

Figure 4: *From top to bottom*: sequence of temporal observation fd; sequence of red predominance observation h with unsupervised parameter estimation ($H_{lip} = 130$; $\Delta_H = 6$); sequence of temporal observation thresholded with unsupervised parameter estimation ($\theta_{fd} = 9$); sequence of red prevailing observation thresholded ($\theta_h = 192$).

3. THE SEGMENTATION ALGORITHM

3.1. The spatiotemporal MRF framework

From these two thresholded observations, four initial labels (a_0,a_1,b_0,b_1) are derived, for coding four pixel classes: pixels with ($_1$) (resp. without ($_0$)) motion, belonging (a) (resp. not belonging (b)) to red hue areas. This label field is supposed to follow the main MRF (Markov Random Field) property related to a *spatiotemporal neighborhood* structure, *i.e.* the label l_s of the current pixell s depends only on the labels of its spatiotemporal neighbors n.

Maximizing the A Posteriori probability (MAP criterion) of the label field is equivalent to minimizing a global energy function [2]:

$$W(S) = \sum_{o\in\{fd,h\}} U_o(S) + \alpha.U_m(S) \quad (13)$$

where U_o and U_m represent respectively the *attachment energies* (expressing the link between labels and observations, Eq. 14) and the *model energy* (corresponding to spatial and temporal a priori constraints) (Eq. 15) over the image S, α is a weighting coefficient between the two energies.

$$U_o(S) = \sum_{s\in S} \left[\frac{[o_s - \psi_o(l_s)]^2}{2\sigma_o^2} \right] \quad (14)$$

where ψ_o is an attachment function, mean value of the observation o over S and σ_o^2 is the corresponding variance. Both are estimated on line.

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

$$U_m(S) = \sum_{s\in S} \Big[\sum_{n\in\eta(s)} V_{st}(l_n, l_s) \Big] \quad (15)$$

The spatiotemporal potential function V_{st} is defined as the inverse of the Euclidian distance between two neighbors. The distance integrates two elementary potentials β_s and β_t as scale factors (Eq. 16).

$$V_{st}(l_n, l_s) = \frac{\beta_s(l_n,l_s)\beta_t(l_n,l_s)}{\sqrt{\beta_t(l_n,l_s)^2\left(\delta_x^2 + 4\delta_y^2\right) + \beta_s(l_n,l_s)^2\delta_t^2}} \quad (16)$$

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

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 (details in [4]).

3.2. The relaxation algorithm

The iterative deterministic algorithm ICM (Iterated Conditional Modes) is implemented to compute the minimum energy at each site (Eq. 13 with typ. $\alpha = 20$), starting from the thresholded fields as initial label configuration. After a few iterations on the field (less than 10 to respect the stopping criterion for convergence $\Delta W(S)/W(S) < 0.05$ %), convergence is achieved. One obtains homogeneous red hue and lip motion fields (Fig. 5).

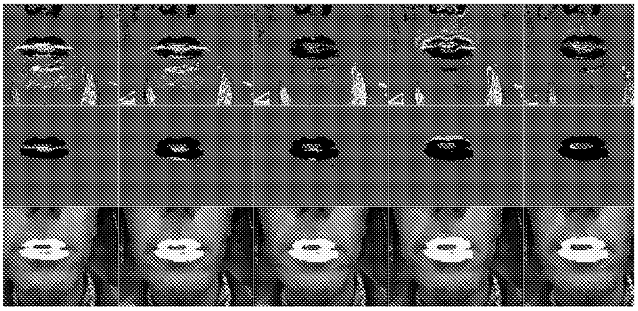

Figure 5: *From top to bottom*: initial labels; label fields after relaxation, *The 4 labels are shown in grey levels (from white to black:* b_1, a_1, b_0, a_0*)*; red hue relevant label images (a_0 and a_1) superposed with luminance.

3.3. ROI estimation

From lip red hue relevant labels, the ROI is evaluated *on line* by maximising a cost function $\Gamma(S)$ on each image (Eq. 10 in [4]) after each step of the relaxation. One each image, the last estimated ROI is increased with a scale factor and used to initialize the current one. The ROI estimation reduces the relaxation time by surrounding the mouth precisely. Moreover, it increases the accuracy of parameter's estimation.

4. LIP SHAPE EXTRACTION

Different sequences have been tested, some with natural red make-up (*Top* in Fig. 6), others with poor lighting conditions without any make-up (*Bottom* in Fig. 6). These results show the robustness of the unsupervised algorithm to the variability of natural conditions. The unsupervised parameter estimation method gives a one pixel mean difference with ground truth measures for the vertical height and the horizontal width of the internal lip opening (Fig. 7). The external shape is unfortunately more elusive but accurate enough to initiate a simple deformable geometrical model.

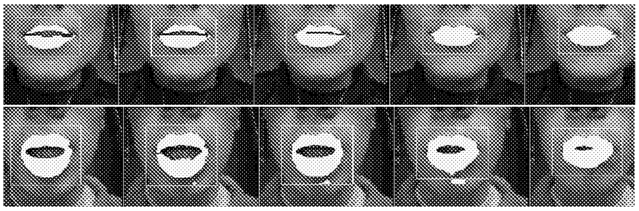

Figure 6: *Top*: Sequence of final red hue fields with ROI superposed on the luminance with soft red make-up; *Bottom*: Sequence of final red hue fields with ROI superposed on the luminance with no lighting supply and no make-up.

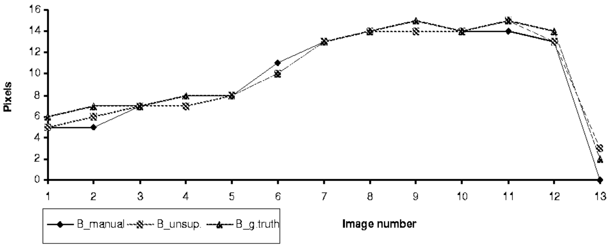

Figure 7: Internal lip measurement on the sequence partially shown Fig. 5 (B_{manual} as manual parameter setting measure; $B_{unsup.}$ as unsupervised parameter estimation measure; $B_{g.truth}$ as ground truth measure)

5. CONCLUSION

An unsupervised lip segmentation have been successfully applied to several sequences in natural conditions (natural images of speaker's face without any particular make-up or lighting). First, the choice of a logarithmic transformation close to the characteristic of the human visual system enables the algorithm to estimate accurately the mean value of lip hue H_{lip} (the speaker's dependant parameter). This transformation is combined with a noise estimation on the frame difference. Second, the spatiotemporal algorithm integrates hue with motion information, improving the quality of contours often elusive on speaker's lips. Finally, the quality of the segmented fields is similar to those obtained with parameters determined beforehand manually [4]. We need to process more sequences to test the robustness of the parameter estimation with more difficult cases, like faces with beard or colored people faces.

The proposed algorithm requires less than 10 iterations until convergence (about 2 sec. on a SunUltra1).

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] 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.

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

[4] M. Liévin and F. Luthon. Lip features automatic extraction. In *Proc. of the* 5^{th} *IEEE Int. Conf. on Image Processing*, Chicago, Illinois, October 1998. (accepted paper).

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

[6] T. Wark and S. Sridharan. A synthetic approach to automatic lip feature extraction for speaker identification. In *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pages 3693–3696, Seattle, Washington, USA, May 1997.