Lip Motion Automatic Detection

F. Luthon, M. Liévin
Image Processing and Pattern Recognition Laboratory,
Grenoble National Polytechnical Institute,
LTIRF, 46 av. Félix-Viallet, 38031 Grenoble Cedex, France
email : luthon@tirf.inpg.fr fax : +33 4 76 57 47 90

Abstract

An algorithm for speaker's lip motion detection is presented, based on the processing of a colour video sequence of speaker's face under natural lighting conditions and without any particular make-up. It is intended for applications in speech recognition, videoconferencing or speaker's face synthesis and animation.
The algorithm is based on a statistical approach using Markov Random Field (MRF) modelling, with a spatiotemporal neighbourhood of the pixels in the image sequence. Two kinds of observations are used : the temporal difference between successive images (motion information) and the purity of red hue in the current and past images (spatial information about lip location). The field of hidden labels, relevant for lip motion detection, is obtained by energy minimisation and proves to be robust to lighting conditions (shadows).
This label field is used to extract qualitative information (mouth opening and closing) but also quantitative information by measuring some geometrical features (horizontal and vertical lip spacing) directly on the label field.

Key words : motion detection, MRF, lip motion, automatic speech recognition.

1 Introduction

It is well known that human beings use visual information as a precious help to enhance their understanding of speech, especially in noisy conditions. Looking at the speaker's face greatly improves audio recognition ability. So the processing of video information, in addition to the processing of audio signal, is very interesting for automatic systems aiming at recognizing or synthesising speech and speaker faces.

Recently, many approaches have been proposed in this area, based on colour analysis (*e.g.* Vogt in [7]), dynamic contours (*e.g.* Dalton *et al.* in [7]), deformable templates (*e.g.* Silsbee in [7]), kroma key [1]... Some of them require special blue make-up, and/or special lighting conditions, hardly achievable for practical applications.
Here we propose an original algorithm for lip motion detection that works in natural conditions, the only requirement being that the camera is fixed w.r.t. the speaker's head, and that illumination conditions are quasi-constant. In the final application, a micro-camera with a microphone will be mounted on a light helmet worn by the speaker.

The outline of the processing (and the paper organisation) is as follows: first, *RGB* colour space is transformed into *HIP* (*Hue, Intensity, Purity*) colour space. Observations taken in this space are described in section 2. Then, some hidden labels with their interaction energies, relevant for lip motion detection, are defined in an MRF modelling framework (section 3). A deterministic relaxation algorithm is used to maximise the a posteriori probability of label field w.r.t. observations. Finally, lip parameters extraction is performed directly on the label field after relaxation, and qualitative information about mouth opening and closing is also obtained (section 4).

2 Observations in *HIP* colour space

First, an *RGB* image sequence of mouth movements is acquired at full video-rate (25 images/s, 8 bits/pixel colour). The region of interest spans from nostrils to chin and contains the lip region. The speaker needs no particular make-up and the video system no particular illumination. But the camera position is supposed to be static w.r.t. speaker's head.

To obtain robust observations from image sequences, we transform the *RGB* colour space into the *HIS* (*Hue*, *Intensity*, *Saturation*) colour space:

$$H = \frac{\pi}{2} - \arctan\left(\frac{2R-G-B}{\sqrt{3}(G-B)}\right) + k \quad (1)$$

$$I = \frac{R+G+B}{3} \quad (2)$$

$$S = 1 - \frac{\min(R,G,B)}{I} \quad (3)$$

where: $k = 0$ if $G > B$ and $k = \pi$ otherwise[1].

Indeed, grey-level techniques are known to be sensitive to light conditions: shadows make the border of the lips to be elusive. Colour techniques are less sensitive to lighting variations. Since red colour usually prevails in lips, we develop a technique to extract *close-to-red-hue prevailing regions*. This allows to be less dependent on colour variations (yellow light for example). So, there is no need for specific illumination or make-up of lips.

On the other hand, techniques based on luminance differences (*e.g.* spatial or temporal gradients) are accurate to detect motion at pixel level, but often fail to get lip regions.

Therefore, it seems interesting to have a mixed processing of luminance and hue components. Unfortunately, saturation proves to be not enough discriminating. Instead of it, we use *colour purity* P as defined in [6]:

$$P = I.S = \frac{R+G+B}{3} - \min(R,G,B) \quad (4)$$

This quantity is more suited to extract shadows from images. P is high for pure colours and near zero for shadow areas. Indeed, shadows correspond to almost equal values of the three components R, G, B of a pixel. So we work in the *HIP* colour space, where the three components prove to be better decorrelated. Our purpose is to localise lip regions with red hue purity, and to detect motion with luminance differences. Under the hypothesis of *static camera* and *quasi constant illumination*, motion detection is related to temporal changes of the intensity function (luminance). Lip localisation is related to red hue prevailing regions.

Based on the *HIP* colour space, two kinds of observations are derived. The *temporal observation* at a site or pixel $s = (x, y)$ at time t is simply the frame difference $fd(s)$ computed on the luminance between successive images. The *spatial observation* $h(s)$ is computed from the hue and purity, assuming a Gaussian repartition of lip hue:

$$fd(s) = I_t(s) - I_{t-1}(s) \quad (5)$$

$$h(s) = \left[256 - \left(\frac{H(s) - H_m}{\sigma}\right)^2\right] \times 1_{P(s)>\delta} \times 1_{|H(s)-H_m| \leq 16.\sigma} \quad (6)$$

where $H(s)$, $I(s)$ and $P(s)$ are respectively the hue, intensity (or luminance) and purity at pixel s, H_m is the mean value of lip hue (determined beforehand on the first image), σ is the standard deviation of lip hue (heuristic value, typ. $4 \leq \sigma \leq 9$) and δ is an heuristic threshold on purity (typ. $50 \leq \delta \leq 100$). Whenever needed, a temporal index t is added in the notations (*e.g.* $I_t(s)$). The notation $1_{condition}$ denotes a binary function which takes the value 1 is the condition is true, 0 otherwise.

Fig. 1 illustrates the observations h on a typical sequence, and Fig. 2 shows the corresponding observations fd on the same sequence.

The lip-MAD (Motion Automatic Detection) algorithm uses three observations at a time: $h_t(s)$, $h_{t-1}(s)$ and $fd(s)$. These observations are thresholded (thresholds γ and θ) and yield three low-level informations at site s (basic coding on 3 digits): two about the presence of red hue at times t and $t-1$ (r_t, r_{t-1}) and one about motion sign (m) (see Table 1). Fig. 3 illustrates these initial information fields (marked in upper case letters, *e.g.* R_t). Note the improvement obtained by introducing a threshold δ on purity (the teeth visible in Fig. 3c disappear in Fig. 3d). But these initial fields are noisy and have to be regularized for robustness. For that purpose, the statistical framework of MRF modelling is used, as explained in the next section.

3 Motion Labels in MRF Framework

The combination of these three information fields M, R_t and R_{t-1}, each information at a site s taking two or three different values, contributes to define an initial label field $\hat{L}_t$, made of 12 distinct *labels* l_i where $l \in \{a, b, c, d\}$ and

[1] In practice, a modified formula avoiding the use of *arctan* is used for computing H [5].

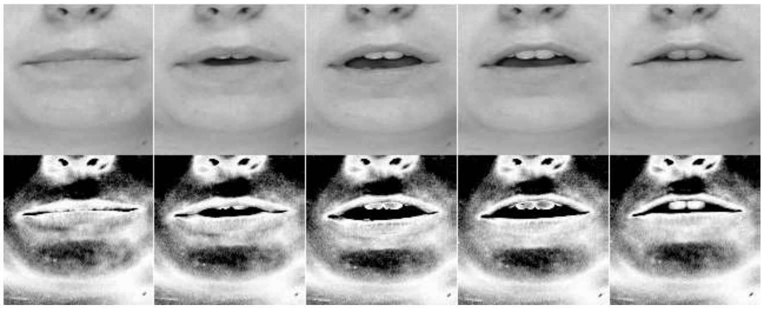

Figure 1: *From top to bottom* : sequence of luminance images; sequence of spatial observations h_t at five time instants t (red-hue prevailing regions in white, $\delta = 0$).

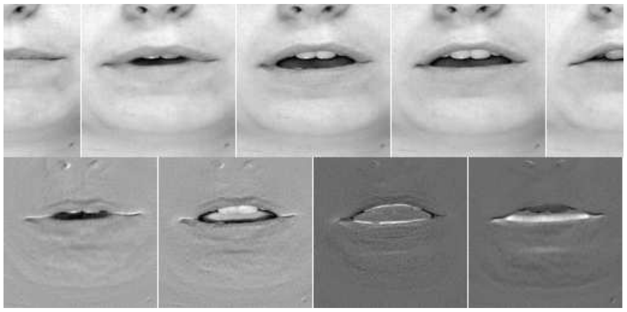

Figure 2: *From top to bottom* : same sequence of luminance images; sequence of temporal observations fd (positive values in white, negative in black, zero in grey).

$i \in \{0, 1, 2\}$ (see Table 1).

This label field is supposed to follow the main MRF property related to *spatiotemporal neighbourhood* structure η shown on Fig. 4: *i.e.* the label $l(s)$ of a pixel s depends only on the labels of its neighbours $n \in \eta(s)$, not on the whole image. Using the MAP criterion (Maximum A Posteriori) and the equivalence between MRFs and Gibbs distributions, it is easily shown that maximizing the a posteriori probability of the label field w.r.t. observations is equivalent to minimizing a global energy function U [3]. In our case, U is made of five terms:

$$U = \sum_{s \in S} \left[\lambda [U_{h_t}(s) + U_{h_{t-1}}(s)] + U_{fd}(s) + U_{sp}(s) + U_{tp}(s) \right] \tag{7}$$

Table 1: Observations, low-level informations and the 12 corresponding initial labels (typ. $\theta = 10$, $\gamma = 100$).

observations			initial detection at site s		coding			initial labels
$h_t(s)$	$h_{t-1}(s)$	$fd(s)$	red hue	motion sign	$r_t(s)$	$r_{t-1}(s)$	$m(s)$	$l(s)$
$< \gamma$	$< \gamma$	$\lvert . \rvert < \theta$	$\emptyset$	$\emptyset$	0	0	0	a_0
		$> \theta$		$+$			1	a_1
		$< -\theta$		$-$			2	a_2
	$> \gamma$	$\lvert . \rvert < \theta$	$t-1$	$\emptyset$		1	0	b_0
		$> \theta$		$+$			1	b_1
		$< -\theta$		$-$			2	b_2
$> \gamma$	$< \gamma$	$\lvert . \rvert < \theta$	t	$\emptyset$	1	0	0	c_0
		$> \theta$		$+$			1	c_1
		$< -\theta$		$-$			2	c_2
	$> \gamma$	$\lvert . \rvert < \theta$	t & $t-1$	$\emptyset$		1	0	d_0
		$> \theta$		$+$			1	d_1
		$< -\theta$		$-$			2	d_2

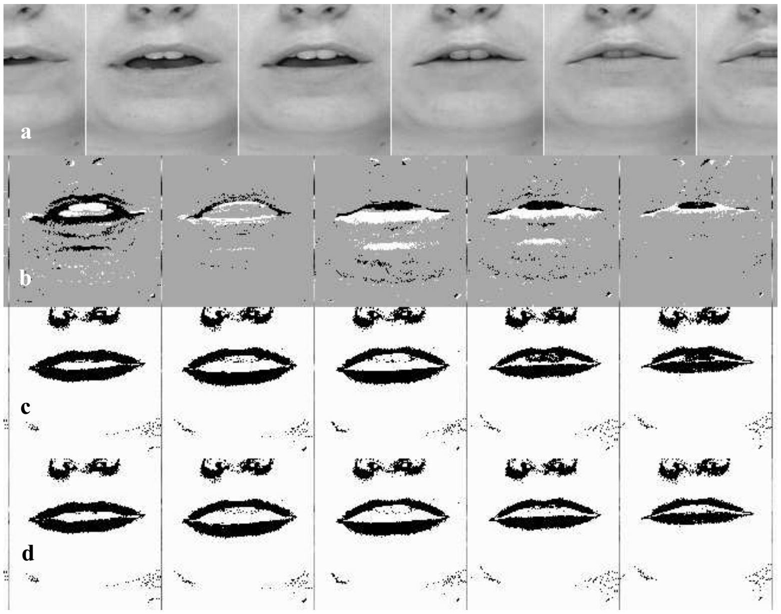

Figure 3: *a)Image sequence; b)Initial motion fields M (grey $= 0$; white $= 1$; black $= 2$); c)Initial red hue fields R_{t-1} with $\delta = 0$; d)Initial red hue fields R_{t-1} with $\delta = 100$.*

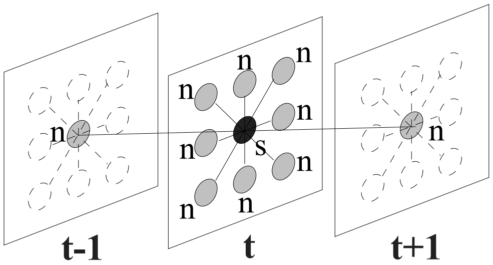

Figure 4: Spatiotemporal neighbourhood structure η with binary cliques $c = (s, n)$. s is the current pixel (in black), n is any spatiotemporal neighbour of s (in grey).

where S is the current image and λ is a weight factor (typ. $\lambda = 4$).

The first three terms $U_{h_t}, U_{h_{t-1}}$ and U_{fd} of Eq.(7) are three *adequation energies* (expressing the link between labels $l(s)$ and the three observations at site s) which are of the very classical form:

$$U_o(s) = \frac{[o(s) - \psi_o(l(s))]^2}{2\sigma_o^2} \tag{8}$$

where observation o is either h_t, h_{t-1} or fd. σ_o^2 is the corresponding observation variance, which may be estimated on line or empirically fixed. ψ_o is an adequation function defined in Table 2: the values taken be ψ_o are simply the mean values of corresponding non zero observations.

The last two terms of Eq.(7) are two *a priori energies* (associated to spatial and temporal a priori modelling) ensuring the regularization (robustness) of the final label field after relaxation (*i.e.* after energy minimization). They are defined as sums of potential functions:

$$U_{sp}(s) = \sum_{n \in \eta_{sp}(s)} V_{sp}(s, n) \tag{9}$$

$$U_{tp}(s) = \sum_{n \in \eta_{tp}(s)} V_{tp}(s, n) \tag{10}$$

where V_{sp} (resp. V_{tp}) are *potential functions* associated to spatial (resp. temporal) cliques. Only *binary cliques*

Table 2: Computation of ψ_o (in third column, $l \in \{a, b, c, d\}$ and $i \in \{0, 1, 2\}$).

ψ_o	observation $o(s)$	label $l(s)$	observation set S_o	$\psi_o(l(s))$
ψ_{fd}	$fd(s)$	l_0	$\{s \in S/\lvert fd(s)\rvert < \theta\}$	0
		l_1	$\{s \in S/fd(s) > \theta\}$	$\frac{1}{N_o}\sum_{s\in S_o} o(s)$ with $N_o = card(S_o)$
		l_2	$\{s \in S/fd(s) < -\theta\}$	
ψ_{h_t}	$h_t(s)$	a_i, b_i	$\{s \in S/h_t(s) < \gamma\}$	
		c_i, d_i	$\{s \in S/h_t(s) > \gamma\}$	
$\psi_{h_{t-1}}$	$h_{t-1}(s)$	a_i, c_i	$\{s \in S/h_{t-1}(s) < \gamma\}$	
		b_i, d_i	$\{s \in S/h_{t-1}(s) > \gamma\}$	

$c = (s, n)$, made of current pixel s and one of its neighbours n, are considered in the neighbourhood $\eta(s)$. $\eta_{sp}(s)$ (resp. $\eta_{tp}(s)$) $\subset \eta(s)$ is the subset of spatial (resp. temporal) neighbours of current site s.

These potential functions represent interaction energies between s and n: V_{sp} tends to reinforce spatial homogeneity, while V_{tp} tends to emphasize temporal variations (fast motion of lips). They are made of elementary potentials defined in Table 3:

$$\begin{aligned} V_{sp}(s,n) &= \left[V_m(s,n) + V_{r_t}(s,n) + V_{r_{t-1}}(s,n) + V_{m,r}(s,n) + V_{r,m}(s,n)\right]\alpha(s,n) \\ V_{tp}(s,n) &= W_m(s,n) + W_{r_t}(s,n) + W_{r_{t-1}}(s,n) \end{aligned} \qquad (11)$$

where $\alpha(s, n)$ is a weighting coefficient taking into account the orientation of spatial cliques $c = (s, n)$ (with $n \in \eta_{sp}(s)$):

- $\alpha(s, n) = 1$ for vertical cliques.
- $\alpha(s, n) = 4$ for diagonal cliques.
- $\alpha(s, n) = 8$ for horizontal cliques.

This anisotropy increases label field homogeneity in horizontal directions (higher interaction energies).

Table 3: Definition of elementary potentials V and W (typ. $\beta_0 = 400, \beta_2 = 2000, \beta_1/\beta_2 = 32$). All permutations between s and n and between t and $t-1$ hold, yielding all required values in Eq.(11).

<table>
<tr><th>pixel s</th><th colspan="2">neighbour n</th><th>spatial potentials</th><th>temporal potentials</th></tr>
<tr><td colspan="3">m</td><td>V_m</td><td>W_m</td></tr>
<tr><td colspan="3">same values</td><td>$-\beta_2$</td><td>$+\beta_0$</td></tr>
<tr><td>1 or 2</td><td colspan="2">0</td><td>$+\beta_2/2$</td><td>$-\beta_0/2$</td></tr>
<tr><td>2</td><td colspan="2">1</td><td>$+\beta_2$</td><td>$-\beta_0$</td></tr>
<tr><td colspan="3">r</td><td>V_r</td><td>W_r</td></tr>
<tr><td colspan="3">same values</td><td>$-\beta_1$</td><td>$-\beta_0/2$</td></tr>
<tr><td>1</td><td colspan="2">0</td><td>$+\beta_1$</td><td>$+\beta_0/2$</td></tr>
<tr><td>m</td><td>r_{t-1}</td><td>r_t</td><td>$V_{m,r}$</td><td></td></tr>
<tr><td rowspan="2">1 or 2</td><td>0</td><td>1</td><td>$-\beta_2$</td><td></td></tr>
<tr><td colspan="2">other</td><td>$+\beta_2$</td><td></td></tr>
<tr><td>0</td><td colspan="2">any</td><td>0</td><td></td></tr>
</table>

An iterative deterministic relaxation algorithm (ICM : Iterated Conditional Modes) [2] is implemented to compute the minimum energy at each site, starting from the initial label configuration $\hat{L}_t$ derived from Table 1. Because of the spatiotemporal neighbourhood chosen (Fig. 4), the *past and future* label fields are required for computing U_{tp}. The past is the result of the previous relaxation at time $t - 1$. For the future, we compute a coarse initialisation $\hat{L}_{t+1}$ in the same way as we get $\hat{L}_t$. Note that the requirement for having the future at disposal implies a *one image delay* in the lip-MAD algorithm.

After a few iterations (typ. 5) on the image, convergence is achieved and the noise that was present in the initialization fields disappears after relaxation (compare Fig. 3 and 5). One obtains a robust field of 12 labels (Fig. 5c), but only a subset of these labels is relevant for extracting lip motion. Indeed, only the four labels b_1, b_2, c_1, c_2 corresponding to an effective displacement ($fd \neq 0$) and to the presence of lips at a single time t or $t-1$ are significant of lip shape modification and useful for motion detection and mouth parameters measurement, as presented in the next section.

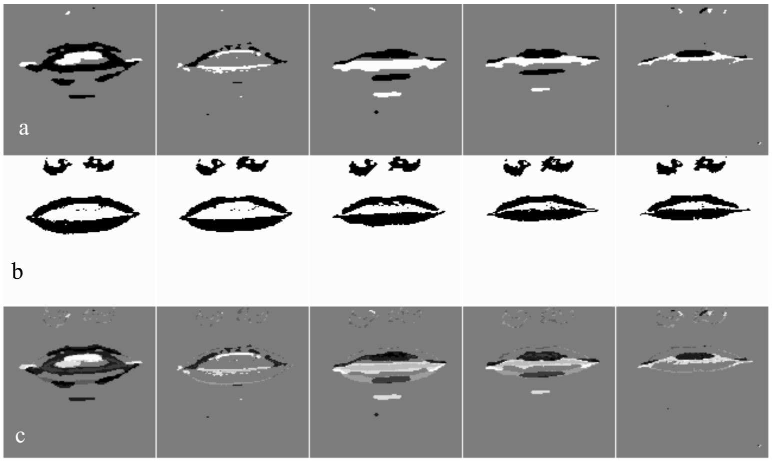

Figure 5: Fields after relaxation. *a)Final motion fields M; b)Final red hue fields R_t with $\delta = 100$; c)Final label fields L_t: the 12 labels are shown in grey levels.*

4 Mouth Parameters and Motion Extraction

The lip-MAD algorithm aims at extracting *dynamically* (*i.e.* in time) various geometrical parameters measured on speaker's face corresponding to mouth main features (Fig. 6): vertical height B and horizontal width A of internal lip contour; vertical height B' and horizontal width A' of external mouth borders; lip opening surface S. One can assess the quality of the measures by verifying that $A.B$ is strongly correlated to S [1].

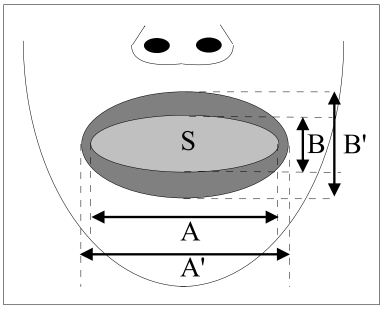

Figure 6: Mouth parameters measured on speaker's face.

To get the useful motion information and mouth parameters, one first needs a coarse localization of mouth region. For that purpose, we use two functions f_{γ_Y} and f_{γ_X} based on the computation of vertical spatial gradients of H and on mouth shape a priori constraints [4]:

$$\forall y / s \in S_1 : \qquad f_{\gamma_Y}(y) = \sum_x \left| \frac{\partial H(s)}{\partial y} \right| \times \left(1 - \left(\frac{x - x_0}{x_0} \right)^2 \right) \tag{12}$$

$$\forall x : \qquad f_{\gamma_X}(x) = \sum_{y / s \in S_2} \left| \frac{\partial H(s)}{\partial y} \right| \tag{13}$$

where x_0 is the horizontal coordinate of the image center, and S_1 is a subset of the image (two thirds of the vertical dimension) centered vertically (this eliminates the contribution of nostrils and chin). The function $f_{\gamma_Y}(y)$ accounts for high jumps in vertical gradients, indicating the presence of lips, and is weighted quadratically along x in order to focus on lip shape center. This function is computed and then thresholded (threshold γ), yielding a first vertical bounding box S_2: $S_2 = \{s = (x, y) \in S / y_{min} \leq y \leq y_{max}\}$.

Once S_2 is obtained, function $f_{\gamma_X}(x)$ is computed and also thresholded (same threshold γ), yielding the final bounding box S_3 including the mouth area: $S_3 = \{s = (x, y) \in S / y_{min} \leq y \leq y_{max}$ and $x_{min} \leq x \leq x_{max}\}$. The coordinate y_0 of the corner of lips is given by the maximum of $f_{\gamma_Y}(y)$. Fig. 7 illustrates the behaviour of functions f_{γ_Y} and f_{γ_X} on a typical image and gives the bounding box detected in five consecutive images.

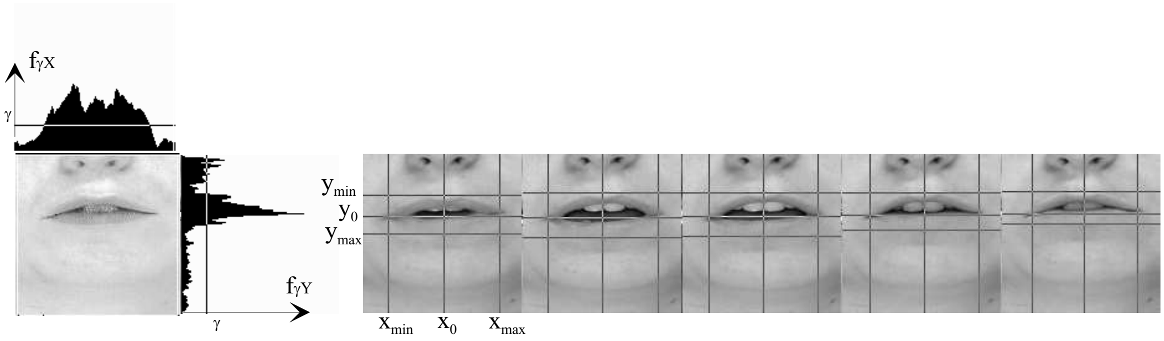

Figure 7: Gradient functions f_{γ_Y} and f_{γ_X} computed on H and thresholded ($\gamma = 100$) to get mouth bounding box.

Then, we extract, count and track in this bounding box the specific labels associated to each tracked lip border. Indeed, it is easily shown that, depending on illumination conditions, a specific lip border in motion corresponds to only two specific labels (*e.g.* b_2, c_1 or b_1, c_2). An automatic extraction and counting of these specific labels (heuristic methods detailed in [4]) gives the nature and amplitude of mouth movements (dynamic features extraction, see Fig. 8).

Of course, static measurements (*i.e.* in space only) of parameters may also be performed directly on the hue information field R_t [4], since it gives clean masks of the lips with good accuracy (as shown in Fig. 5b). This allows an automatic static extraction of face parameters, *e.g.* by edge detection on field R_t (results not shown here).

But our experience is that dynamic features extracted from label field L_t (integrating temporal behaviour of lips) yield richer information than simple frame by frame static measurements on field R_t (better precision in lip tracking before and after a phoneme is produced, good detection of complete closing of mouth, better anticipation of lip motion in case of prephonatory movement). A typical example of result in natural lighting conditions is given in Fig. 8. Note that the algorithm does not fail when mouth is completely closed.

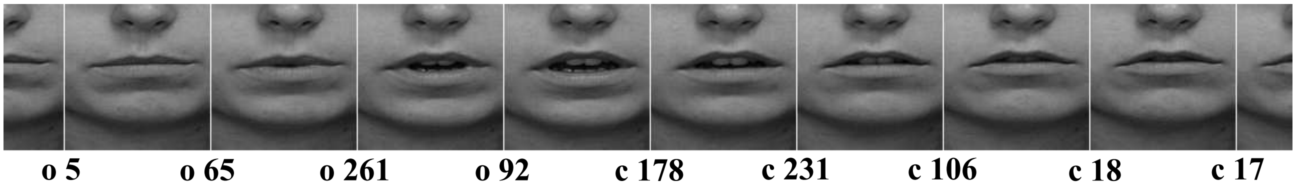

Figure 8: Automatic detection of mouth opening (o) and closing (c) with quantitative values giving the amplitude of motion.

5 Conclusion

As a result of this early work on lip motion automatic detection using MRF modelling, we would like to put emphasis on a few points, that we believe to be very important for processing video sequences of speaker faces. First, the use of colour instead of grey-level images enhances greatly the achievable results. The key point is then to find the good transformation from *RGB* to another colour space (*e.g.* *HIP*) for the problem at hand. In our case, colour purity was better suited than colour saturation.

Secondly, integrating temporal information with spatial information, *i.e.* processing together a few images consecutive in time, instead of processing each image independently, is a key point for improving motion analysis of face features (*e.g.* lips).
Thirdly, in the case of a video sequence acquired in natural lighting conditions, some kind of regularisation (*e.g.* via MRF modelling) is required after collecting relevant low-level observations on speaker's video sequence, as long as contours in a face are often elusive, due to noise or shadows.
The proposed algorithm for lip motion detection requires further investigation: we have to process many more video sequences to assess the robustness of our method. We also have to study some more difficult cases like faces with beard, or coloured people faces. But the first results shown here are quite promising.

Acknowledgments

We thank C. Benoît and his team from Speech Communication Institute (ICP, Grenoble), for fruitful collaboration and discussions.

References

[1] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of French visemes for visual speech synthesis. In G. Bailly, C. Benoît, and T.R. Sawallis, editors, *Talking Machines: Theories, Models and Designs*, pages 485–504, Amsterdam, North-Holland, 1992. Elsevier Science Publishers B.V.

[2] J. Besag. On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B*, 48(3):259–302, 1986.

[3] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on PAMI*, 6(6):721–741, November 1984.

[4] M. Liévin. Détection du mouvement des lèvres par analyse dynamique de séquences d'images. Master's thesis, Nat. Polytech. Institute, Grenoble, France, September 1996. Option Signal-Image-Speech.

[5] H. Palus and D. Bereska. The comparison between transformations from RGB colour space to IHS colour space, used for object recognition. In *5th IEE International Conference on Image Processing and its Applications*, pages 825–827, Heriot-Watt University, Edinburgh, UK, July 1995.

[6] A.J. Pritchard, R.E.N. Horne, and S.J. Sangwine. Achieving brigthness-insensitive measurements of colour saturation for use in object recognition. In *5th IEE International Conference on Image Processing and its Applications*, pages 791–795, Heriot-Watt University, Edinburgh, UK, July 1995.

[7] D.G. Stork and M.E. Hennecke, editors. *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series F: Computer and Systems Sciences.* Springer-Verlag, Berlin, 1996.