Active Contours and Logarithmic Hue-like Colour Space Applied to Lip Tracking

Patrice Delmas
Department of Computer Science, University of Auckland
Private Box 92019, Auckland, New Zealand
patrice@cs.auckland.ac.nz

Marc Lievin
Surgical Systems Laboratory, Caesar Research Center
Bonn, Germany
http://www.caesar.de/ssl

Abstract

Extracting speech intelligibility meaningful face features is one of the most actively researched fields in Image Processing. So far, no system is able to precisely extract contours of lips, eyebrows, eyes in changing conditions. This paper focuses on external and internal lip contours extraction and introduces some variations on active contours theory as well as a broader development of the skin-oriented hue-like colour space. Extracted contours on diverse lip shapes under varying illumination conditions are then provided.

Keywords: active contours, colour space, lip contours,

1 Introduction

Over the years and since the first research work published on lip and face feature extractions [1], a tremendous number of publications have been concentrating on locating, tracking faces and subsequently on estimating face features contours. Indeed, it is commonly acknowledged that the speech visual information, mainly the mouth area, is an essential component of the speech intelligibility process which can help the listener under degraded acoustical conditions [2]. Furthermore, new developments in Human-Computer Interface have proven that a realistic synthetic face facilitates the interaction with the human user. Recently, results dealing with stereo-cameras system and providing 3D lip contours have been published [3]. Still, realistic (i.e. looking like the real one), finely detailed lip contours are scarcely obtained in the published works. Although they might not be mandatory for audiovisual speech processing, where the in-between lip area seems to carry enough visual information [4], they are necessary in applications such as synthetic talking faces (videoconference, interactive agent, and cartoon animation), user verification and recognition (audiovisual biometric features). Other applications for audio-visual interface include communication for disabled people and MPEG4 transmission

Our hierarchical face feature extraction algorithm first segments the face in several regions (usually the skin, lip and inner lip areas) using hue, motion and clustering observations embedded in a Markov Random Field relaxation method. The mouth area being located, characteristics points of the mouth (namely commissures and vertical extrema) are extracted. Finally, using gradient and region information from the hue image, active contours converge towards the lip contours. This paper will not focus on the face location and face features tracking over video-sequences components, already published by the authors in [5] and [6], but on the definition of the hue-like colour space and on some extension of the active contours, both essential to accurate lip contours extraction. The next chapter will introduce the theory behind the logarithmic hue-like colour space (now called LUX). The third chapter will briefly review the active contours theory and introduce two important notions: the string snake and a hue-based statistic for inner snake initialisation. The following chapters will then provide some results and conclude this paper.

2 The LUX space

The LIP (Logarithmic Image Processing) model is a mathematical framework which was introduced in [7], for the explicit purpose of grey-level image representation and processing. This model was later extended by [6] in order to facilitate the representation of colour images and specifically engineered for the purpose of real-time skin segmentation. The model, known as the LUX colour space, has been demonstrated to provide some robustness to changing illumination and also emphasizes hue in regions characterized predominantly by either the red - as is the case for human skin [6] - or blue channel.

The LUX colour space consists of the components (L, U, X) which correspond respectively to the notions of luminance, hue and saturation.

$$L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \qquad (1)$$

$$U = \begin{cases} \frac{M}{2}(\frac{R+1}{L+1}) & if\ R < L, \\ M - \frac{M}{2}(\frac{L+1}{R+1}) & otherwise. \end{cases} \qquad (2)$$

$$X = \begin{cases} \frac{M}{2}(\frac{B+1}{L+1}) & if\ B < L, \\ M - \frac{M}{2}(\frac{L+1}{B+1}) & otherwise. \end{cases} \qquad (3)$$

Where M corresponds to the dynamic range of grey-levels i.e. 256 for 8-bit coding, and R, G and B correspond to the components of the RGB space.

The LIP model is defined in the continuous case by three equations: a transform f from the intensity space (variable x) to the space of tones (variable y), an isomorphism ϕ from the space of tones onto a logarithmic space (variable $\overline{x}$) and an inverse isomorphism ϕ^{-1} [6].

$$f : x \to y = f(x) = M\Big(1 - \frac{x}{x_0}\Big) \qquad (4)$$

$$\phi : y \to \overline{x} = \phi(y) = -M \ln\Big(1 - \frac{y}{M}\Big) \qquad (5)$$

$$\phi^{-1} : \overline{x} \to y = \phi^{-1}(\overline{x}) = M(1 - \exp^{-\frac{\overline{x}}{M}}) \qquad (6)$$

where $x \in]0, ..., x_0]$ is a continuous grey level and $x_0 \in]0, ..., M]$ is the maximum transmitted light.

The isomorphism, $\Phi = \phi \circ f$, provides a logarithmic transform normalised by x_0.

$$\Phi : x \to \overline{x} = M \ln\Big(\frac{x_0}{x}\Big) \qquad (7)$$

$$\Phi^{-1} : \overline{x} \to x = x_0 \exp\Big(-\frac{\overline{x}}{M}\Big) \qquad (8)$$

As the components $R, G, B \in [0, M[\times[0, M[\times[0, M[$ in the discrete case, rgb are taken as $r = R + 1$, $g = G + 1$ and $b = B + 1$ in order to maintain the interval $]0, M]$ as is required by the LIP theory [6]. The transformed variables lux are noted similarly i.e. $l = L + 1$ etc.

The following illustrates the construction of the LUX colour space:

$$\begin{tikzcd}
(R,G,B) \arrow[r, "+1"] \arrow[d] & (r,g,b) \arrow[r, "\Phi"] \arrow[d, "\Psi"] & (\overline{r},\overline{g},\overline{b}) \arrow[d, "T"] \\
(L,U,X) & (l,u,x) \arrow[l] & (\overline{l},\overline{u},\overline{x}) \arrow[l, "\Phi^{-1}"']
\end{tikzcd}$$

The isomorphism, Φ, transforms the $[r, g, b]$ vector into its logarithmic counterpart $[\overline{r}, \overline{g}, \overline{b}]$.

$$\begin{cases} \overline{r} = M \ln \frac{r_0}{r} \\ \overline{g} = M \ln \frac{g_0}{g} \\ \overline{b} = M \ln \frac{b_0}{b} \end{cases} \qquad (9)$$

The transform T is then applied to $[\overline{r}, \overline{g}, \overline{b}]$, yielding the vector $[\overline{l}, \overline{u}, \overline{x}]$:

$$\begin{cases} \overline{l} & = 0.3\overline{r} + 0.6\overline{g} + 0.1\overline{b} \\ \overline{u} & = \overline{r} - \overline{l} \\ \overline{x} & = \overline{b} - \overline{l} \end{cases} \qquad (10)$$

If we let r_0, g_0 and b_0 correspond to the maximal values of the colour channels r, g and b respectively and combine Eq's. 9 and 10 then,

$$\begin{aligned} \overline{l} &= 0.3M \ln \frac{r_0}{r} + 0.6M \ln \frac{g_0}{g} + 0.1 \ln \frac{b_0}{b} \\ &\Rightarrow M \left[\ln\left(\frac{r_0}{r}\right)^{0.3} + \ln\left(\frac{g_0}{g}\right)^{0.6} + \ln\left(\frac{b_0}{b}\right)^{0.1} \right] \\ &\Rightarrow M \ln \left(\frac{r_0^{0.3} g_0^{0.6} b_0^{0.1}}{r^{0.3} g^{0.6} b^{0.1}} \right) \qquad (11) \end{aligned}$$

Also, by letting l_0, u_0 and x_0 correspond to the maximal values of l, u and x, and using Eq's. 8, 10 and 9 we yield:

$$l = l_0 \frac{1}{r_0^{0.3} g_0^{0.6} b_0^{0.1}} r^{0.3} g^{0.6} b^{0.1} \qquad (12)$$

$$u = u_0 \left(\frac{g_0^{0.6} b_0^{0.1}}{r_0^{0.7}} \right) \left(\frac{r^{0.7}}{g^{0.6} b^{0.1}} \right) \qquad (13)$$

$$x = x_0 \left(\frac{g_0^{0.6} r_0^{0.3}}{b_0^{0.9}} \right) \left(\frac{b^{0.9}}{g^{0.6} r^{0.3}} \right) \qquad (14)$$

At this point two assumptions are made:

1. the components r_0, g_0, b_0 are close to the maximal intensity I_0. This assumption can be justified given that, the camera has been white balanced, calibrated for the full range of white values.

2. the maximal luminance L_0 is close to the dynamic range M. additionally, the relation $l_0 = u_0 = x_0 = M$ is imposed to ensure that an equivalent dynamic range is maintained.

This reduces to Eq's. 15 - 19:

$$l = r^{0.3} g^{0.6} b^{0.1} \tag{15}$$

$$\begin{aligned} u_+ &= \overline{r} - \overline{l} \\ &= M\frac{r}{l} \quad where\, \overline{r} \geq \overline{l} \Rightarrow R < L \qquad (16) \\ x_+ &= \overline{b} - \overline{l} \\ &= M\frac{b}{l} \quad where\, \overline{b} > \overline{l} \Rightarrow B < L \qquad (17) \\ u_- &= \overline{l} - \overline{r} \\ &= M\frac{l}{r} \quad where\, \overline{r} < \overline{l} \Rightarrow R > L \qquad (18) \\ x_- &= \overline{l} - \overline{b} \\ &= M\frac{l}{b} \quad where\, \overline{b} < \overline{l} \Rightarrow B > L \qquad (19) \end{aligned}$$

Under the *strong* assumption that skin is predominantly characterized by the red channel i.e. $R > L \Leftrightarrow \overline{r} < \overline{l}$ it is sufficient to take only the contribution of $\overline{u}_-$, and as such, we may define the red chroma as:

$$U = \begin{cases} M\frac{L+1}{R+1} & if\ R > L, \\ M - 1 & otherwise. \end{cases} \tag{20}$$

Noting, from Eq. 1, that the luminance formula L within the LUX space is simply a weighted geometric mean of the R,G and B components - it can be inferred that regardless of the image, the difference between the luminance and Y or G channels will be minimal. Thus Eq. 20 can be further simplified to:

$$\widehat{U} = \begin{cases} 256 \times \frac{G}{R} & if\ R > G, \\ 255 & otherwise. \end{cases} \tag{21}$$

The ratio G/R is scaled by the multiplicative constant in order to adjust its range to the 8-bit quantization levels (M = 256)

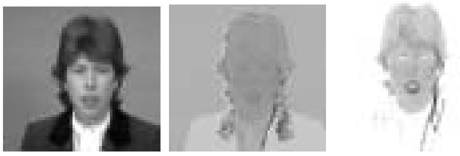

Figure 1: From left to right: an image extracted from the Claire sequence, the corresponding angular hue and the logarithmic hue-like image obtained using the LUX space.

3 Active contours

Over the years and since the first research work on lip and face feature extractions [1], methods have been divided in global and local approaches. Local approaches are based on grey-level or colour values of a pixel or a group of pixels, gradient and sometimes spatio-temporal information ([6]). Statistical values over a region or projection [8] of the pixels information can be used to assess whether they belong to face features regions. Still these methods apart from [6] are often inconsistent over faces as they heavily rely on heuristic considerations, such as the number of peaks and valleys of the curve obtained by horizontal projection and concatenation of the image pixels. However, when applied locally, i.e. when the location of the face features is roughly known, they can help simplify the feature contour extraction. Global approaches usually rely on shape constraints and local information (grey level, colour or gradient values) of the image. The main approaches are Deformable Templates, Active Shape Models and Active Contours. Deformable templates introduced by [9] rely on simplistic description of face features and tend more to locate the features rather than delineating their precise contours. The first deformable template lip model was composed of three quadrics, two parabolas and had up to ten parameters accounting for the position, angle, width, length and respective heights of the lip contours. The latest deformable template describes the face contour, nose, lip, eyes and eyebrows contours through about one hundred control points [10]. Still it usually requires a rough manual fitting to the face and rarely provides contours detailed enough to retain speech intelligibility information. Active shape models have been introduced by [11]. They use the statistical analysis (usually via principal component analysis) of the lip shapes variations through locution to define a generic deformable model of lip contours and its mode of variations. The first step is the classification of the face feature contours database. The technique has been developed by [12] for face analysis and consists on the computation of the eigenvectors of the autocorrelation matrix, formed by the concatenation of the vectors describing the face features (usually via a set of points). The contours are then supposed to evolve only through shapes that can be described as a mean (of the database set of features) contour and a linear combination of the largest eigenvectors (that can be seen as contours), their associated weighting coefficients defining their respective influence. Although the process provides a way to control the contour shapes, there is no clear indication on how the database set composition influences the allowed shape variations. Furthermore, the database set of face contours, that should be as large as possible to account for all the likely face feature shapes, is usually determined manually. This combined with the expression-less aspect of the extracted contours refrain from using that method as long as automatic and precise method extracting face feature contours does not exist. Active contours introduced by [13] are an energy-based method which evolves through the minimization of their functional which is a balanced combination between internal constraints (based on bending and stretching physical properties of thin plates) and external constraints (based on image information which describe the features to extract). They were primarily introduced as an interactive method where the user could help the evolving process being attracted (via the adjunction of additional energetic term know as string) to some interesting part of the image or instead pushed (via energetic term known as volcano) away from other specific regions. For their ability to adapt to various situations without exhaustive testing, parameters setting or database management, the active contour technique has been chosen to extract lip contours.

3.1 Theory

An active contour is a constrained chained contour (usually defined by a set of points v (eq. 22) evolving through the minimisation of its associated energy functional Φ (eq. 23). The active contour moves towards its final position while constantly balancing the respective influence of its internal energy (eq. 24) and external energy, also known as perturbation.

$$v(s) = [x(s), y(s)]\,,\ s\epsilon[0,1] \qquad (22)$$

$$\Phi : v(s) \longrightarrow \int_0^1 \left(E_{int}(s) + E_{ext}(s)\right) ds \qquad (23)$$

The internal energy (Eq. 24) is a second order regularization term derived from the Tikhonov ill–posed problem theory [14] which controls the curve bending and stretching properties via the parameters α and β. The external energy (Eq. 25) has its minima near the image features to extract. Active contours achieve best when they are set up to extract image contours, i.e. when their external energy has its minima on edges. To do so, a Gaussian filtered edge map of the hue image is used (eq. 25):

$$E_{int}(s) = \alpha|v'(s)|^2 + \beta|v''(s)|^2 \qquad (24)$$

$$E_{ext}(s) = -\left|\nabla\left(G_\sigma \otimes H\right)(v(s))\right|^2 \qquad (25)$$

with ∇ represents the gradient operator, G_σ the 2D Gaussian kernel and H the hue image. This leads to the classical dynamic scheme (Eq. 26) where I_d is the identity matrix, A the Toeplitz snake matrix, V the snake control points vector and $\frac{1}{\gamma}$ the time step coefficient.

$$V(t) = \left(I_d + \frac{A}{\gamma}\right)^{-1} \left(V(t-1) - \frac{1}{\gamma}F(V(t-1))\right) \qquad (26)$$

where F represents forces derived from external energy ($F_{ext}(v) = -\nabla(E_{ext}(v(s)))$).

The matrix $I_d + \frac{A}{\gamma}$ is called the snake stiffness matrix and is a narrow band (of width 5) quasi-pentadiagonal Toepliz circulant matrix [13]. Its inverse is usually approximated using the LU inversion technique, which can be computationally expensive for large matrices. Under a few assumptions, which can be found in [5] it has been demonstrated that the general element m_l of the matrix inverse is equal to:

$$m'_l = \frac{1}{N}\sum_{k=1}^{N} \frac{cos\left(\frac{2(l-1)(k-1)\pi}{N}\right)}{1 + \frac{\frac{4}{h^4}\sin^2\frac{k\pi}{N}(4\beta\sin^2\frac{k\pi}{N} + h^2\alpha)}{\gamma}} \qquad (27)$$

with $(I + \frac{A}{\gamma})^{-1}$, being, for an even number of control points (our case in this paper), a symmetric circulant matrix with its first line of the following form:

$$\begin{bmatrix} m'_1 & m'_2 & \ddots & m'_{N/2} & m'_{N/2+1} & m'_{N/2} & \ddots & m'_2 \end{bmatrix} \qquad (28)$$

3.2 String snake

Although not discussed in this paper, the corners (or commissures) of the mouth precise location is an essential step towards accurate lip contour extraction. Previous work trying to extract mouth features without commissures detection lead to failure [15]. Using the extracted corners as anchor points for an active contour has been proven successful [5]. Presented here is an improvement of this technique where the active contour is anchored to the corner points via strings (fig. 2), thus providing more freedom of movement and therefore helping obtaining better results. For an N points active contours, the snaxels V_0 and $V_{\frac{N}{2}}$, are then connected to the commissures via strings. The active contour evolution equation then becomes:

$$V(t) = \left(I_d + \frac{A}{\gamma}\right)^{-1} \left(V(t-1) - \frac{1}{\gamma}F(V(t-1))\right) \qquad (29)$$

$$+F_{string}(fix_l, V_0) + F_{string}(fix_r, V_{\frac{N}{2}})) \qquad (30)$$

where fix_l and fix_r are respectively the left and right position of the commissures points (see fig. 2). A string-like link between $M_i(v_i)$ and $M_j(v_j)$, creates at point M_i a force equal to:

$$F_{string}(v_i, v_j) = -\left(\frac{\partial E}{\partial v}\right)_i \quad (31)$$

$$F_{string}(v_i, v_j) = 2k\overrightarrow{M_iM_j} \quad (32)$$

$$E_{string}(v_i, v_j) = -k\|v_i - v_j\|^2 \quad (33)$$

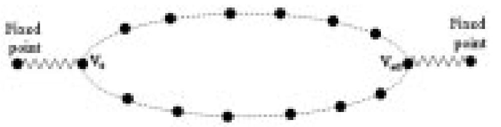

Figure 2: Closed active contour anchored to fixed points via strings.

3.3 hue based statistics for inner snake initialisation

The major problem encountered while extracting lip contours is, the outer-lower lip contour detection and the initialisation of the inner lip active contour. The outer-upper lip contour is always well detected thanks to a clear edge transition between the skin and the upper-lip. It has to be pointed out as well that outside the lip corner area (where shadows annihilate image information), the hue, derived from the LUX space, is a good estimator of the lip region. Furthermore, both lips usually have the same colour characteristics. Therefore, a lip colour statistic can be derived from the study of the pixels situated:

- on a few lines under the upper-outer lip contour (as the upper lip is thin)
- on 2/3 of the columns under the lip boundary (to avoid the colour-less corners area)

The mean (m_{lip}) and variance (σ_{lip}) of the hue over the previously defined area is computed. By supposing that the distribution of hue values over the lip area follows a Gaussian law, the probability a pixel $p(x, y)$ belongs to the lip area is given by $P_{m,\sigma}(p(x, y)) \leq \eta$ with:

$$P_{m,\sigma}(p(x, y)) = \frac{1}{\sqrt{2}\sigma} \exp -\frac{{p(x, y) - m_{lip}}^{2}}{2\sigma^2} \quad (34)$$

The inner active contour points are then initialised on the first pixels marked as "not belonging to the lip region", encountered inwards the outer active contours. Associated with the string-snake, this usually provides accurate lip contours.

4 Results

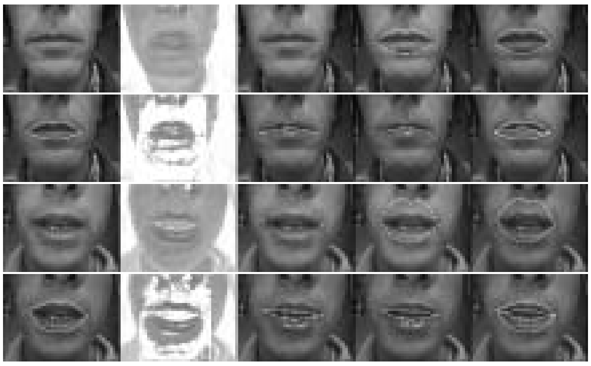

Figure 3: Detected borders on two different speakers (top: Benny, bottom: Aktham). From the plain image to the lip contours extraction: from left to right; from top to bottom : Luminance image, hue image, mouth commissures, initial snake points, corresponding spline, external snake after convergence, statistics on lip area, initial internal snake points obtained thereafter, corresponding spline, internal and external snakes after convergence.

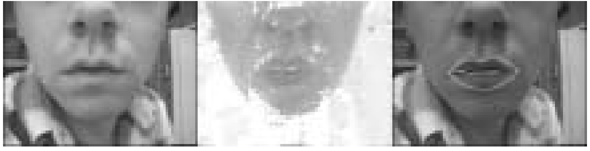

Figure 4: Detected borders on Marc using the Labiophone helmet.

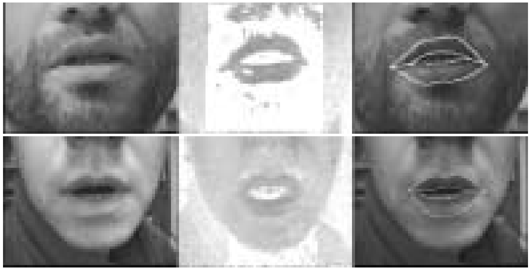

Figure 5: Detected borders on Patrice (top) and Marc (bottom) using a Sony EVI-D100 video camera.

Several results are presented here, encompassing various speakers with diverse lip shapes either showing a closed or open mouth. Some of the images have been acquired via a camera-mounted w.r.t the head, others have been obtained with the widely used Sony EVI-D100 camera. Finally, some images (fig. 6) have been acquired under asymmetric lighting conditions. All the parameters (namely α, β, γ and η) have been set up equal to the same values for all the

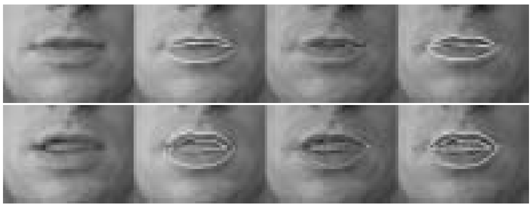

Figure 6: From speech to image: from left to right: Lip area image, contours manually extracted, extraction of the lip contours by the by the presented approach, superposed results.

experiments to account for the robustness of the method. For the last results provided ((fig. 6), second images), the contours have been delineated by a speech processing 'expert' for sake of comparison. Although the maximum vertical distance between both (manual and obtained with our algorithm) contours never exceed five pixels, it has been found inadequate for speech-from-image applications. Furthermore, a detailed study of the lip contour manually extracted by a speech-processing specialist has proven that the provided contours do not always rely on image processing information such as edges, clustered areas, etc.

5 Conclusion

In this paper, the theory behind the logarithmic hue-like colour space has been developed. Some improvements of the active contours theory, applied to lip contours extraction, have enhanced the robustness of the results. Furthermore, comparison between the results expected by speech processing 'expert' and our results has partially proven the inadequacy of image processing based methods for audio-visual recognition applications. However, it does not jeopardize the relevance of the proposed method for applications such as 3D face synthesis and low-bit rate coding videoconferencing. Currently, developments towards 3D lip contours extraction (via stereovision), 3D lip synthesis using a 3D generic model and derivation of the active contour stiffness inverse matrix formula for open snakes are underway.

References

[1] E.D. Petajan, B. Bischoff, D. Bodoff, and N.M. Brooke. An improved automatic lipreading system to enhance speech recognition. *CHI88*, pages 19–25, 1988.

[2] B. LeGoff, T. Guiard-Marigny, M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibility of talking faces. *2nd ETRW on Speech Synthesis*, pages 53–56, 1994.

[3] G. Loy, R. Goecke, S. Rougeaux, and A. Zelinsky. Stereo 3D lip tracking. *International Conference on Control, Automation, Robotics and Computer Vision*, December 2000.

[4] L. Reveret. *Conception and evaluation of an automatic system for labial gestures tracking.* PhD thesis (in french), Institut National Polytechnique de Grenoble, 1999.

[5] P. Delmas. *Lip Contour Extraction by Means of Active Contours. Application to multimodal communication.* PhD thesis (in french), National Polytechnic Institute of Grenoble, 2000.

[6] M. Liévin. *Entropico-logarithmic analysis of colour video-sequences applied to segmentations and tracking of speakers face.* PhD thesis (in french), Institut National Polytechnique de Grenoble, 2000.

[7] 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.

[8] P. Radeva and E. Marti. Facial features segmentation by model-based snakes. *International Conference on Computer Analysis of Images and Patterns*, 1995.

[9] A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.

[10] Z. Xue, S.Z. Li, D. Shen, and E.K. Teoh. A novel bayesian shape model for facial feature extraction. *International Conference on Control, Automation, Robotics and Computer Vision*, December 2002.

[11] T.F. Cootes, T.J. Taylor, D.H. Cooper, and J. Graham. Active shape models: Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[13] M. Kass, A.Witkin, and D. Terzopoulos. Snakes: Active contours models. *International Journal of Computer Vision*, pages 321–331, 1988.

[14] A. Tikhonov and V. Arsenine. *Méthodes de résolution de problèmes mal posés.* MIR, 1974.

[15] B. Leroy. *Deformable templates applied to face recognition.* PhD thesis (in french), Université Paris IX-Dauphine, June 1996.