Towards Robust Lip Tracking

Patrice Delmas
CITR, University of Auckland
patrice@cs.auckland.ac.nz

Nicolas Everse
LIS-INPG, Grenoble
eveno@lis.inpg.fr

Marc Lievin
Research Center Caesar, Bonn
http://www.caesar.de/ssl

Abstract

An algorithm for the automatic extraction of speakers' lips in video sequences is presented here. Our goal is to extract minimum face feature parameters, vital for audio-visual communication, in adverse conditions and at a very low bit rate coding. Our method uses spatial (region and contour) and temporal (similarity function) information from Luminance and Hue components. A new literal inverse of the active contours stiffness matrix is introduced. This ensures a fast and accurate convergence of active contours towards lips boundaries. The use of Kanade-Lucas tracking algorithm with our point extraction method leads to an automatic, fast and robust initialization of Snakes. The significant robustness enhancement and related computational cost decrease allow processing approaching real time.

1. Introduction

It is commonly observed that visual information provides precious help to the listener or can interfere (McGurk effect, cocktail-party effect) with the perceived sound under degraded acoustical conditions [5]. The goal of the present work is to extract visual information for automatic speech recognition (ASR), audio-conferencing and speaker’s face synthesis under natural lighting conditions. Some approaches proposed in this area are based on gray level analysis (e.g. Luettin in [7]). Others use color analysis, but need to determine optimal values of some parameters (e.g. Cianiz in [7]). Works on real time lip-tracking [8] usually rely on manual location of features or fine tuning of several parameters. Others will focus on inner (e.g. Petajan in [7]) or outer lip contour extraction, rarely both. Here, an algorithm is proposed for inner and outer lip contour tracking under natural conditions for video-sequences. It integrates a presegmentation phase to precisely extract characteristic points of the mouth [3] (see figure 3) on the first image of the sequence. Then, lip contour are detected and tracked through video sequence using active contours [3] combined with a variant of the Kanade-Lucas feature points tracking algorithm, adapted to the particular geometry of our desired features. The requirement is 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/pixel/pixel) contains the region of the face spanning from chin to nostrils. Furthermore, "real time" processing rate is mandatory to maintain a sufficient speech intelligibility [5].

2. Lips contours extraction and tracking

2.1 Extraction of characteristic points

The precise location of mouth characteristic points (commisures and vertical extrema of the mouth) is essential to ensure accurate contour extraction methods [7]. A method has been developed to accurately find corners
(see figure 3) and vertical extrema of the mouth, also known as characteristic points (see figure 2). Further details regarding this part of our algorithm are detailed in [3].

Figure 2. Lip corners for different speakers.

Figure 3. (a): quartic and parabolas based mouth model; (b): Characteristic points (CP) of the lips extracted.

2.2 Characteristic points tracking

The process described in section 2 allowed the extraction of characteristic points of the mouth in the first frame of a video sequence. Their positions in the following frames are obtained using a variant of the Kanade–Lucas algorithm [9] adapted to the particular geometry of the features considered. Here, the studied space is reduced to the closed neighborhood of the characteristic points, which allows a significant gain in processing time compared to the direct extraction method (section 2.1). The neighborhoods of the points being tracked are assumed to have only translation movements from image $I$ to next image $J$. The position of one characteristic point in image $J$ can be expressed as:

$$ J(x, y) = I(x - \alpha, y - \beta) + n(x, y) $$

(1)

Where $n(x, y)$ is the noise level for the pixel $(x, y)$ and $\alpha, \beta$ is the displacement vector. $I(x, y)$ and $J(x, y)$ are scalars, for example, the luminance value of the pixel $(x, y)$. Figure 4 illustrates equation 1 for a mono-dimensional signal. The vector $d$ is chosen to minimize the residue factor $\epsilon$, computed on the neighborhood window $W$, around the pixel $(x, y)$, as follow:

$$ \epsilon = \int \int_W [I(x - d) - J(x)]^2 w(x) \, dx $$

(2)

Where $x = (x, y)^T$ and $w(x)$ is a weighting function usually constant equal to 1. Resolution of equation 2 (detailed in [9]) leads to the 2x2 linear system of equations:

$$ Gd = e $$

(3)

Resolution of this equation gives an estimation of vector $d$ and therefore of the characteristic points location in image $J$. During the tracking process, the neighborhoods of characteristic points were supposed to only move through translations. More realistic models (affine models) taking into account deformation movements have been proposed [9].

However, for consecutive images of a video sequence, the window $W$ deformation is small. Therefore, using affine models considerably slows down the tracking process without significant gain in the accuracy of the results. Furthermore, the interior of the mouth is a highly deformable area in which similarity calculation from one image to the next is very difficult to achieve. Teeth and tongue can appear or disappear very quickly during aperture or closure. Analysis windows are thus shifted in a way they does not encroach upon the interior of the mouth (see figure 5). Some results of tracking without shifted windows are shown in figure 6.

Figure 4. A neighborhood in image $J$ may be found in the next image $I$ applying a translation of vector $d$.

Figure 5. lips-enabled neighborhood used to track characteristic points (cross). Interior points and commissures searching windows (dashed windows) are shifted to avoid mouth interior area.

Figure 7 shows results obtained from a video sequence using shifted analysis windows of size 21x21 pixels. The Kanade-Lucas almost allows the correct tracking of the characteristic points (figure 7 left side). However, if lips movements are fast, shadowing effects may happen and the
interior points can be difficult to follow (see figure 7 left side, last image). For each point, the residue error $e$ gives the accuracy estimation. When the residue error is too important, the point is initialized using the method briefly described in section 2.1 and detailed in [3]. As this is only performed on mispredicted points (usually one or two) this method is faster than the one used to initialize the first image. The model is then adjusted deforming only one or two parabola.

2.3 Lip contours extraction

Usually, the contour extraction methods combine edge detection (global to the image) to a chaining process (often based on local information in the image). Due to their ability to integrate these two steps in a single process and their adaptability to the desired features, active contours has been chosen to extract lip boundaries. They are defined as chained contours evolving, from a predefined initial form, through the minimization (often of a gradient descent type) of an energy functional using image data (a gray level, gradient-based or distance transform map) related to the desired features. Introduced by Kass et al [4] active contours were designed for interactive interpretation in which the user guides (by external forces modification) the snake near the desired solution. A snake is a parameterized curve $v$ defined (Eq. 5) by its Cartesian coordinates $x$ and $y$ along the curvilinear abscissa $s$, which evolves through the minimization of its functional $\Phi$ (Eq. 6):

\[
v(s) = \begin{bmatrix} x(s), y(s) \end{bmatrix}, \quad s \in [0,1]
\]

\[
\Phi : v(s) \rightarrow \int_0^1 (E_{int}(s) + E_{ext}(s)) ds
\]

The internal energy (Eq. 7) is a second order regularization term derived from Tikhonov ill-posed problems theory. It controls the curve smoothness via weighting parameters $\alpha$ and $\beta$. $\alpha$ controls the snake tension and $\beta$ its curvature. External energy (Eq. 8) represents the fitting of image data to the current vector. The image gradient to extract edge points. To do so a classical gradient filter (such as Sobel or Canny-Deriche) is used:

\[
E_{int}(s) = \alpha |v'(s)|^2 + \beta |v''(s)|^2
\]

\[
E_{ext}(s) = -\nabla (G_{\sigma} \otimes I)(v(s))^2
\]

$\nabla$ represents the gradient operator, $G_{\sigma}$ the 2D Gaussian kernel and $I$ the current image. This leads us to the classical dynamic scheme:

\[
V(t) = (A + \gamma I_d)^{-1} (\gamma V(t-1) - F(V(t-1)))
\]

where $I_d$ is the identity matrix, $A$ the Toeplitz snake matrix, $V$ is the snake control points vector, $F$ the force derived from external energy and $\gamma$ the time step coefficient. The matrix $I_d + \frac{A}{\gamma}$ is called the stiffness matrix and is a narrow band (of width 5) quasi–pentadiagonal Toeplitz circulant matrix [4]. Its inverse is usually approximated using the LU inversion technique, which can be computationally expensive for large matrices. A new way to compute the exact inverse of this matrix is introduced here. Previous work
[1] gave the eigenvalues of the stiffness matrix. Under a few assumptions which can be found in [3] it has been demonstrated that the general element $m_l$ of the stiffness matrix inverse is equal to:

$$m_l' = \frac{1}{N} \sum_{k=1}^{N} \frac{\cos \left( \frac{2(l-1)(k-1)\pi}{N} \right)}{1 + \frac{\delta^2}{\gamma^2} \sin^2 \left( \frac{4\delta \sin^2 \frac{\delta}{2} + \beta^2\gamma}{\gamma} \right)}$$

(10)

With $N$ number of snake's control points and $h$ distance between two consecutive control points. The inverse snake stiffness matrix, $(I + \frac{A}{\gamma})^{-1}$, for an even number of control points, our case in this paper, is of the following form:

$$
\begin{pmatrix}
    m_1' & m_2' & \cdots & m_{N/2}' & m_{N/2+1}' & \cdots & m_N'
    \\
    m_2' & m_1' & \cdots & m_{N/2-1}' & m_{N/2}' & \cdots & m_N'
    \\
    \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots
    \\
    m_{N/2+1}' & \cdots & \cdots & m_1' & m_2' & \cdots & m_{N/2}'
    \\
    \vdots & \vdots & \ddots & m_{N/2-1}' & m_{N/2}' & \cdots & m_N'
    \\
    m_N' & \cdots & \cdots & m_{N/2-1}' & m_{N/2}' & \cdots & m_1'
\end{pmatrix}
\left(\begin{array}{cccc}
    m_1' & m_2' & \cdots & m_{N/2}' & m_{N/2+1}' & \cdots & m_N'
    \\
    m_2' & m_1' & \cdots & m_{N/2-1}' & m_{N/2}' & \cdots & m_N'
    \\
    \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots
    \\
    m_{N/2+1}' & \cdots & \cdots & m_1' & m_2' & \cdots & m_{N/2}'
    \\
    \vdots & \vdots & \ddots & m_{N/2-1}' & m_{N/2}' & \cdots & m_N'
    \\
    m_N' & \cdots & \cdots & m_{N/2-1}' & m_{N/2}' & \cdots & m_1'
\end{array}\right)
$$

(11)

Active contours are initialized using the shape described in figure 3-a. Its parameters are computed using the characteristic points position for each image. Points of the snake are sampled along these geometrical forms. For the first image, the position of characteristic points is obtained using the method described in section 2.1. For the following frames, their position is computed using our adaptation of the Kanade-Lucas algorithm described in section 2.2. By reducing the number of points needed for the snake initialization, this algorithm allows a significant processing time gain compared to the previous method (section 2.1). This is achieved while maintaining good precision for the estimation of the considered points. Active contours are therefore less sensible to the fitting of its parameters and converge at a faster rate. The new snake stiffness matrix inverse computation [3] and the implementation of the snake algorithm using MMX and SIMD instructions allow processing approaching real time (20 frames per second) for the active contour algorithm. Finally, with no assumption about the lighting conditions, inner and outer lip contours are extracted with accuracy (figure 8) in video sequences.

3. Conclusion

A fast and robust lip-tracking algorithm is presented here. First, corners and vertical extrema of the lips are extracted using Luminance, Hue and gradient information on areas and contours. Next, active contours are used to extract lip contours for the current image. Then a variant of the Kanade-Lucas features point tracking method allows a precise and fast initialization for the following image. Furthermore, it also provides an error criterion that enables a reduction in the number of false initialization for active contours. Current developments go to web-cam approach, cooperation between region and contours, to describe characteristic forms in lip contours, such as commissures or cupidon arc, or obtain precise semantic information (mouth opening or closing) [6]. We are also working on helmet-free acquisitions by locating face and mouth area on more open views using ellipsoid based shapes and region (Hue/Luminance) information [2] or Hierarchical face segmentation based on Markov Random Field relaxation method [6].

Figure 8. Convergence snakes results on two sequences (Top: Nico; Bottom: Benny).

References