15th GRETSI Symposium on Signal and Image Processing, Juan les Pins, 18-22 Sept. 1995, France

ACTIVE CONTOURS FOR LIPREADING - COMBINING SNAKES WITH TEMPLATES

Stefan Horbelt[†] **Jean-Luc Dugelay**

Institute EURÉCOM, 2229 route des Crêtes, B.P. 193, 06904, Sophia Antipolis Cedex, FRANCE
email: {dugelay,horbelt}@eurecom.fr, URL: `http://www.cica.fr/~image`

Nous décrivons dans ce papier un algorithme de suivi d'objets déformables par contours actifs. L'objectif est d'améliorer le taux de reconnaissance en traitement de la parole grâce à des paramètres visuels. Nous proposons une approche conjointe "snake" et "templates déformables" pour initializer et ajuster les paramètres physiques du snake afin d'évaluer avec précision l'état spatio-temporel des lèvres du locuteur (aires, déformations, périmètres). Cette approche conjointe est rendue possible grâce à une connaissance approximative mais a priori du type d'objet à identifier. [§]

keywords: active contours, snakes, template forces, non-rigid motion tracking, bimodal speech recognition

1 Introduction

Current efforts on initialization, modification and combination of snakes with other techniques are described in recent papers [FUJ93, LAI94, TER93, BRE94]. The approach proposed here relies on basic articles about snakes [KAS87], physically-based snakes [SZE91] and deformable templates [YUI92]. It combines basic techniques and extends them for visual speech recognition by lipreading. The initialization of position and parameters is a key problem of snakes [LAI94] as is the inflexibility for deformable templates. The tracking dynamics of the snake is improved by adding mass [SZE91, YUI92], interframe forces or balloon forces [FUJ93]. Models incorporating shape were developed for facial analysis in sequences of images [TER93]. Snakes learning the possible shapes of lips (eigenlips) were designed for lipreading [BRE94].

In the following section (2) we will recall the snake and deformable template models used for lipreading.

Section 3 introduces our snake model which consists of a chain of pole springs, with coil springs between the poles, and nodes with mass. Then we extend the snake with shape by introducing the template force.

This paper proposes an approach appling active contours to track deformable objects. The goal is to enhance speech recognition by including visual parameters. We propose an approach which joins the snake model with the deformable template model to initialize and adjust the physical parameters of the snake in order to evaluate accurately the spatial-temporal state of the speaker's lips (area, deformation, perimeter). This joint approach is possible due to a priori, approximate knowledge about the object's constraints in shape and dynamics.

2 Active Contour Models for Lipreading

A snake [KAS87] is a geometrical curve which approximates image contours through energy minimization. It behaves like an elastic rope that wriggles towards the contour or that slides down the potential hill. The internal forces keep the shape and ensure the spatial and temporal continuity. The external forces pull and guide the snake in an interactive and dynamic iterative process. A snake can be regarded as a mathematical energy minimizing curve or as a physical chain structure driven by forces.

The initialization and normalization of the snake parameters and the local fitting are problematic and require interactive intervention: The snake has to be placed in proximity to the contour and can be trapped in local energy minima. A snake uses only the information from its local surroundings, whereas information about the global shape, like that available to a deformable template, is missing.

A deformable template [YUI92] approximates objects through the variations of its parameters. An appropriate template model must be developed for each object. The unflexible structure and missing dynamics are drawbacks when applied to lipreading. We combine snakes and deformable templates to overcome these shortcomings.

We distinguish among three generations of contour detection: 1. classic *pixel orientated* image processing (e.g. thresholds, morphology, edge filters), 2. active edges like snakes which rely on *local image* information and 3. deformable templates which use *global image* information. Our approach combines all three of these.

Color thresholding, morphology and cluster analysis de-

[†] This work was accomplished as part of the bimodal speech recognition project at the Eurécom Institute, and the Lehrstuhl für Nachrichtentechnik, University Erlangen, Germany.

[§] A french extended summary is available on the world wide web with `ftp://www.cica.fr/pub/papers/SBH95.gretsi_sum_fr.ps.gz`

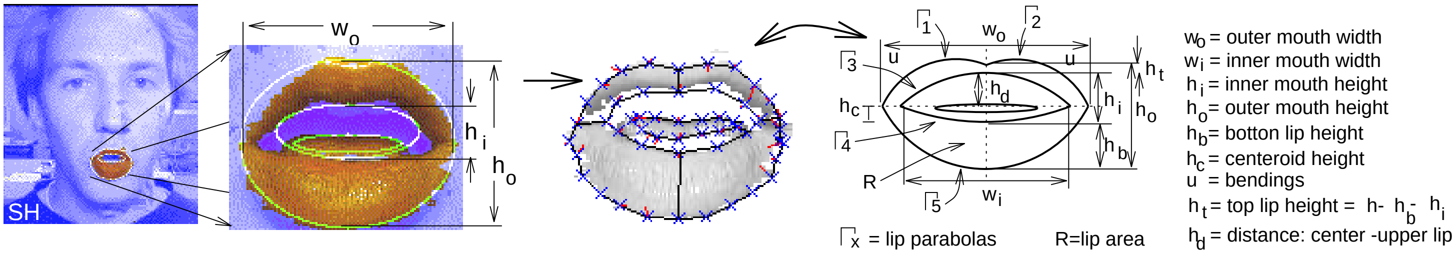

Figure 1: a) Find mouth, b) first parameters, c) lip snake, d) lip template

liver a first approximation of the lip shape (fig.1/a), which is already applicable for speech recognition (fig.3). We use these attributes to parameterize the first ellipsoid template model (fig.1/b) and initialize from it the shape and physical parameters of our lip snake (fig.1/c). Additional template forces will guide the snake. The deformable template acts like an interactive user of the snake. Once the snake has reached a stable state, the deformable template is fitted onto it to verify its shape, to correct its parameters and to add adjusting forces [HOR95].

This loop process may be applied either for each interation or after several iterations of the snake in one image or only once for the initialization of the next image in a sequence of images. The position in the previous image can be used to predict the position in the current image, either with the position and speed of the snake or with the parameters of the deformable template. These three generations of contour detection complete and verify one another.

2.1 Deformable Templates

Several papers describe deformable templates for the mouth [YUI92, HUA92] and apply it to lipreading [HEN94, RAO94]. Yuhas [YUI92] and Rao [RAO94] use parabolas for the lip contours. Huang uses [HUA92] ellipsoids and Hennecke [HEN94] uses parabolas for the inside contours of lips and quadrics for the outside contours.

We propose a modification of the deformable template model of the mouth based on [YUI92]. The model uses parabolas, is symmetrical to the vertical center axis and contains an ellipsoid for the tongue (fig.1/d).

The center point $\vec{v}_c$ of the deformable template is placed at the origin and the rotation of the deformable template is defined by the angle θ (clockwise) between its horizontal axis and the horizontal axis of the picture.

The parabolas for the top of the upper lip (Γ_1 and Γ_2), for the bottom of the upper lip (Γ_3), for the top of the lower lip (Γ_4) and the bottom of the lower lip (Γ_5) are described by the equations:

$$y_{\Gamma_1}(x) = (h_d + h_t)(1 - (\frac{2x}{w_o})^2) - ux - \frac{2u}{w_o}x^2;\ -w_o/2 < x < 0$$

$$y_{\Gamma_2}(x) = (h_d + h_t)(1 - (\frac{2x}{w_o})^2) + ux - \frac{2u}{w_o}x^2;\ 0 < x < w_0/2$$

$$y_{\Gamma_3}(x) = h_d(1 - (\frac{2x}{w_i})^2);\ -w_i/2 < x < w_i/2$$

$$y_{\Gamma_4}(x) = -h_c(1 - (\frac{2x}{w_i})^2);\ -w_i/2 < x < w_i/2$$

$$y_{\Gamma_5}(x) = -(h_c + h_d)(1 - (\frac{2x}{w_o})^2);\ -w_o/2 < x < w_o/2$$

The area of the oral cavity between the upper and the lower lip is $\frac{2}{3}h_i w_i$, the area of the lips is $R = \frac{2}{3}h_o w_o + \frac{uw_o}{12}w_o$.

The parameters for the ellipsoid of the tongue are: y_{ton} = vertical center of tongue, h_{ton} = height of tongue, w_{ton} = width of tongue. The visible area of the tongue is $\frac{\pi}{4}h_{ton}w_{ton}$.

For simplicity we group all parabolas $\{\Gamma_1, \ldots, \Gamma_5\}$ into one contour $\Gamma \in \partial R$ of the deformable template: $\vec{v}(\tilde{s}) = (x, y_\Gamma(x))^\top = (x(\tilde{s}), y(\tilde{s}))^\top$ with the parameterization $\tilde{s} \in [0,1]$ and the discretisation into N nodes $\vec{v}(\tilde{s}(n)), n \in \{1, \ldots, N\}$.

The deformable template tries to minimize its energy. The definition of the energies and the order in which the parameters of the deformable template are changed is crucial for the final result of the energy reduction. Energies can be defined depending on valley (or peak) areas in the image:

$$E_v = \frac{c}{|R|}\int_R \Phi_v(\vec{v})\ dA$$

or on the edges of the image:

$$E_e = \sum_{k=1}^{5} \frac{c_x}{|\Gamma_k|}\int_{\Gamma_k} \Phi_e(\vec{v})\ d\tilde{s}.$$

$\Phi_e(\vec{v})$ and $\Phi_v(\vec{v})$ denote the edge and valley potentials of the image, which are calculated from the raw image [YUI92]. Internal constraint energies keep the deformable template in shape, in our case an elongated shape:

$$E_{con} = \frac{k}{2}(w_o - \lambda h_o)^2;\ \lambda \approx 2.$$

First the area energy function E_v will place ($\vec{v}_c$) and orient the deformable template roughly on the mouth (θ, w_o, h_o). The rest of the parameters are adjusted by the minimization of the edge energy E_e.

The disadvantages of deformable templates are, 1.) a model for each object must be designed, 2.) the energy functions and the rules for how to adjust the parameters of the deformable template have to be created. Thus a deformable template is less flexible in adapting to unexpected forms. Its advantages are low computation costs of the simple analytic model and the small set of parameters describing its shape.

We apply the deformable template of the mouth (fig.1/d) to initialize, to evaluate and to improve the shape of the lip snake model. This will improve the snake incorporating the global shape knowledge of the deformable template. It will also involve a smaller set of parameters for the lipreading.

2.2 Snakes

The basic snake model described by Kass, Witkin and Terzopoulos [KAS87] is a *controlled continuity* spline under the

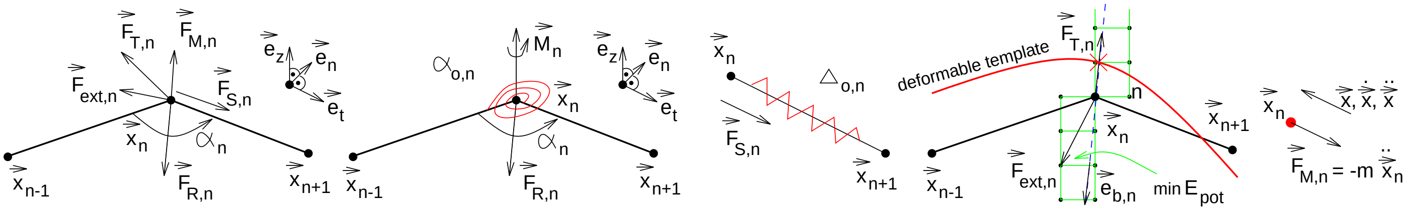

Figure 2: a) 5 forces in node n : b) Rigidity $\vec{F}_{R,n}$, c) Tension $\vec{F}_{S,n}$, d) Template $\vec{F}_{T,n}$, External $\vec{F}_{ext,n}$, e) Mass $\vec{F}_{M,n}$

influence of image forces and external constraint forces. The internal forces keep the snake shape smooth, while the image forces push the snakes toward image features like edges or contours. The external constraint forces are responsible for guiding the snake close to the desired local minimum. These forces can come from a user-interface, automatic attentional mechanisms, or high-level interpretations. Kass gives examples for interactive use only, but we will show how to apply these forces to combine the snakes with the deformable template model.

The energy function of a snake parameterized by $\vec{x}(s) = [x(s), y(s)]^\top$, $s \in [0, 1]$, can be written as

$$E_{snake} = \int_0^1 E_{int}(\vec{x}(s)) + E_{image}(\vec{x}(s)) + E_{con}(\vec{x}(s))\, ds$$

where E_{int} represents the internal forces (tension, rigidity), E_{image} the image forces and E_{con} gives rise to external constraint forces. We did not use the interactive approach, but rather the deformable template approach for developing additional constraint forces, like the template forces. Thus we modify the E_{int} and E_{con} energy terms.

3 Physically-based Active Contour Model Incorporating Shape

We interpret snakes as a chain of pole springs, with coil springs between the poles, and nodes with mass and extend the snake with shape by template forces:

A snake can be seen as a physically-based energy minimizing spline. We parameterize its contour $\vec{x} = [x(s), y(s)]^\top$ with $s \in [0, 1]$ and discretize it into N nodes with $\vec{x}_n = \vec{x}(s(n)), n \in \{1, \ldots, N\}$. Here the active contour is defined as a set of mass nodes, which are connected by nonlinear pole springs and coil springs. It is driven by template forces and external forces. Shape constraints and information are incorporated into the spring parameters at each node. Before we look at the forces we define:

- $\vec{\triangle}_n = \vec{x}_{n+1} - \vec{x}_n$ to denote the difference vector between two connected nodes n and $n + 1$,
- $\triangle_n = ||\vec{\triangle}_n||$ the distance between node n and $n + 1$.

The tripod of orthogonal normal vectors, following the contour, is defined at each node n (fig.2/a):

- $\vec{e}_{t,n} = \vec{\triangle}_n / ||\vec{\triangle}_n||$ denotes the tangential normal vector,
- $\vec{e}_z$ is the binormal vector which points straight outward from the image plane into the reader's direction $\odot$,
- $\vec{e}_{n,n} = \vec{e}_z \times \vec{e}_{t,n}$ is the normal vector.

The bisector angle is defined as $\vec{e}_{b,n} = \frac{\vec{e}_{t,n} - \vec{e}_{t,n-1}}{||\vec{e}_{t,n} - \vec{e}_{t,n-1}||}$.

The snake points are sampled from the deformable template $\vec{x}(s(n)) = \vec{v}(\tilde{s}(n))$. The reference values are initialized with the lip template (Γ). They are defined as:

- the reference angle $\alpha_{0,n} = \cos^{-1} \frac{||\vec{x}_{n+1} - \vec{x}_{n-1}||^2 - \Delta_{n-1}^2 - \Delta_n^2}{2\Delta_{n-1}\Delta_n}$
- the reference distance $\triangle_{0,n} = ||\vec{v}(\tilde{s}(n+1)) - \vec{v}(\tilde{s}(n))||$.

Let us look at the balance of the forces in each node $\vec{x}_n$ in detail (fig. 2/a):

$$\vec{F}_{R,n} + \vec{F}_{S,n} + \vec{F}_{T,n} + \vec{F}_{M,n} + \vec{F}_{ext,n} = 0 \quad \forall \quad \vec{x}_n \qquad (1)$$

1. The rigidity force $\vec{F}_{R,n}$ (fig.2/b) is caused by the torque $\vec{M}_n$ of the coil spring in node n:

   $$\vec{F}_{R,n} = (\vec{e}_{n,n-1}/\triangle_{n-1} + \vec{e}_{n,n}/\triangle_n)||\vec{M}_n|| \qquad (2)$$

   The torque is defined as $\vec{M}_n = \gamma_n \delta_n \vec{e}_z$ with the angle difference $\delta_n = \alpha_n - \alpha_{0,n}$, the reference angle $\alpha_{0,n}$ and the coil spring constant γ_n.

2. The tension force $\vec{F}_{S,n}$ (fig.2/c) is produced by the spring which connects node n and $n + 1$, the reference length $\triangle_{0,n}$ and the pole spring constant c_n:

   $$\vec{F}_{S,n} = c_n(\triangle_n - \triangle_{0,n})\vec{e}_{t,n} \qquad (3)$$

3. The template force $\vec{F}_{T,n}$ (fig.2/d) depends on the distance between the deformable template and the snake node n measured along the bisector angle vector $\vec{e}_{b,n}$:

   $$\vec{F}_{T,n} = \omega_n \vec{e}_{b,n} ||\vec{v}(\tilde{s}(n)) - \vec{x}(s(n))|| \qquad (4)$$

   where ω_n denotes the coupling spring constant. The deformable template $\vec{v}(\tilde{s})$ is parameterized by $\tilde{s} \in [0, 1]$ similar to the snake $\vec{x}(s)$. The template force can include also other a priori constraints imposed by the object with balloon forces [FUJ93] or with forces from springs between other nodes in the snake [TER93].

4. The mass force $\vec{F}_{M,n}$ (fig.2/e) defines the acceleration $\ddot{\vec{x}}_n$ of the node n with the mass m_n:

   $$\vec{F}_{M,n} = -m_n \ddot{\vec{x}}_n \qquad (5)$$

5. The external image force $\vec{F}_{ext,n}$ (fig.2/d) is defined by:

   $$\vec{F}_{ext,n} = -\vec{\nabla} E_{pot}(\vec{x}_n) \qquad (6)$$

   To avoid expensive calculation of the potential energy E_{pot} and its gradient $\vec{\nabla} = (\partial/\partial x, \partial/\partial y)^\top$ for the whole image, a local minimum of the potential energy is searched along the bisector angle vector $\vec{e}_{b,n}$.

3.1 Lip models: Preprocessing and Initialization of Lip Template and Lip Snake

Color thresholding, morphology and cluster analysis allow a first estimation to place the lip template [HOR95].

The snake's nodes $\vec{x}(s(n))$ are placed on sampled points of the deformable template $\vec{v}(\tilde{s}(n))$ and the physical parameters of the snake are initialized. The reference length $\triangle_{0,n}$ and the reference angle $\alpha_{0,n}$ reflect the shape of the deformable template. Assumptions about different dynamics and flexibility of parts of the lips allow the adaptation of the physical parameters in the nodes: The pole spring constant c_n, the coil spring constant γ_n, the coupling spring constant ω_n and the mass m_n.

In the mouth corners and the centers on the top and bottom of the mouth the outer snake is connected with the inner snake by additional springs (fig.1/c) in order to conform the inner snake's shape with the outer contour.

3.2 Discretisation and Prediction in Time

Now all forces can be calculated except $\vec{F}_{M,n}$ which is delivered from the balance of forces (eqn.1, principle d'Alembert: fig.2/d) and we get the acceleration $\ddot{\vec{x}}_n$ (eqn.5). The forward and backward finite difference approximation is applied to compute the snake's velocity $\dot{\vec{x}}_n^{t+\triangle t} = \beta\dot{\vec{x}}_n^t + \ddot{\vec{x}}_n^t \triangle t$, and position $\vec{x}_n^{t+\triangle t} = \vec{x}_n^t + \dot{\vec{x}}_n^{t+\triangle t} \triangle t$, whereas β denotes the velocity damping coefficient (see [TER93]). The prediction in time can be applied during the snake's convergence in one image (intraframe) as well as between images (interframe).

3.3 Join Snakes and Deformable Templates

The deformable template acts like an interactive user who, having global information about possible (lip) shapes, controls and corrects the shape of the snake. The lip template is fitted to the lip snake to get the lip parameters and to adjust the reference parameters $\triangle_{0,n}$ and $\alpha_{0,n}$ of the lip snake. The template forces helps to circumvent local minima in which the snake could be trapped. If the mouth is closed, the inner snake merges to one horizontal line and has difficulties deciding whether it should track the upper or lower lip. To avoid this, we "blow" inside the inner lip snake to produce pressure inside creating a balloon force which presses the snake outside.

4 Conclusion

We described a physically-based active contour incorporating shape applied for lipreading. The introduction of template forces stabilizes the snake's behavior and improves our lip parameters. Preliminarily results on visual speech recognition with DTW are very promising. Our method may be adapated for other applications like cardiac ventricle tracking. Further research is in progress to merge audio and visual speech parameters in a MLP/HMM recognizer. The extracted visual lip parameters can also be applied for lipreading, lip synchronisation, training of hearing impaired people and 3D model-based video coding for videoconferences.

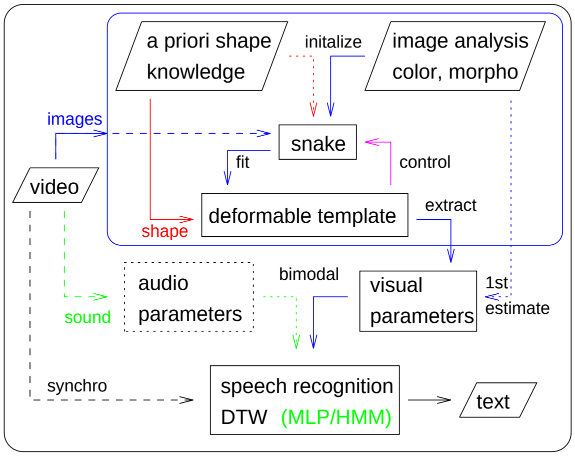

Fig.3.: Data Flow Graph inside Bimodal Speech Recognition System

5 Aknowledgements

The authors thank the researchers at Eurécom and LNT for their useful discussion. This work was accomplished as part of the bimodal speech recognition project at the Eurécom Institute, multimedia communications department, Prof. C.Wellekens, and the Lehrstuhl für Nachrichtentechnik (LNT), Prof. B.Girod, Universität Erlangen-Nuremberg, Cauerstrasse, 91058 Erlangen, Germany.

References

[BRE94] C. Bregler, S. Omohundro, "A Hybrid approach to bimodal speech recognition", *Asimolar 94*, 5p, Nov.94, CA: `http://mambo.ucsc.edu/psl/asilomar.html`.

[FUJ93] K. Fujimura, N. Yokoya, K. Yamamoto, "Motion Tracking of Deformable Objects by Active Contour Models using Multiscale Dynamic Programming", *Jour. of Visual Communication*, 4/4, p382-91, 93.

[HEN94] M.E. Hennecke, K.V. Prasad & D.G. Stork, "Using Deformable Templates to infer visual Speech Dynamics.", *Asimolar 94*, 5p, Nov. 1994, CA

[HOR95] S. Horbelt, "Automatic Lipreading on the Basis of Image Sequences to support Speech Recognition", diploma thesis, Univ. Erlangen, Germany, Apr. 95.

[HUA92] C.-L. Huang, "Human Facial Feature Extraction for Face Interpretation and Recognition", *Pattern Recognition*, Vol.25, No.12, p1435-44, 1992, UK.

[KAS87] M. Kass, A. Witkin, D. Terzopoulos, "Snakes: Active Contour Models", *Int. Journal Computer Vision*, v15/11, p321-31, 93.

[LAI94] K.F. Lai, "Regularization, Formulation and Initialization of the Active Contour Models", Asian Conf. on Comp. Vision, p542-45, 93.

[RAO94] R.R Rao, R.M Mersereau, "Lip modeling for Visual Speech Recognition.", *Asimolar 94*, 5p, Nov. 1994, CA

[SZE91] R. Szeliski, D. Terzopoulos, "Physically-Based and Probabilistic Models for Comp. Vision", *SPIE V.1570 Geometric Methods in Comp. Vision*, p140-52, 91.

[TER93] D. Terzopoulos, K. Waters, "Analysis and Synthesis of Facial Images Using Physical and Anatomical Models", *IEEE Trans. on Pattern Analysis*, p569-579, 93.

[YUI92] A. L. Yuille, P.W. Hallinan, D. S. Cohen, "Feature Extraction from Faces using Deformable Templates", *Int. Jour. of Computer Vision 8:2*, p99-111, 92.

[URL:Lip] Speechreading (lipreading) homepage on WWW: `http://mambo.ucsc.edu/psl/lipr.html`