# EFFICIENT SEARCH METHODS AND DEEP BELIEF NETWORKS WITH PARTICLE FILTERING FOR NON-RIGID TRACKING: APPLICATION TO LIP TRACKING

*Jacinto C. Nascimento* *Gustavo Carneiro**

Instituto de Sistemas e Robótica
Instituto Superior Técnico
1049-001 Lisboa, PORTUGAL

**ABSTRACT**

Pattern recognition methods have become a powerful tool for segmentation in the sense that they are capable of automatically building a segmentation model from training images. However, they present several difficulties, such as requirement of a large set of training data, robustness to imaging conditions not present in the training set, and complexity of the search process. In this paper we tackle the second problem by using a deep belief network learning architecture, and the third problem by resorting to efficient searching algorithms. As an example, we illustrate the performance of the algorithm in lip segmentation and tracking in video sequences. Quantitative comparison using different strategies for the search process are presented. We also compare our approach to a state-of-the-art segmentation and tracking algorithm. The comparison show that our algorithm produces competitive segmentation results and that efficient search strategies reduce ten times the run-complexity.

**Keywords:** Deep belief Networks, lip segmentation, optimization algorithms, search methods, tracking

## 1. INTRODUCTION

It is widely accepted that a key issue to robustness in *Human Computer interface* (HCI) relies on the use of inputs coming from different sensors, offering different modalities of information. An example is the improved performance which can be achieved in *automatic speech recognition* (ASR) systems that use both audio and video input, rather than only audio [1]. The so called *Audio-Video* ASR (AV-ASR) systems have been shown to improve recognition where lip tracking is a crucial and relevant step. Indeed, traditional acoustic based ASR concentrates solely on the acoustic features, where the *Mel-Frequency Cepstral Coefficients* (MFCC) is widely used feature (*e.g.*, [2]). They can perform satisfactorily in a quasi noiseless environment or when its conditions are matched to those presented in the training set. However, these conditions are not always achievable. One strategy to overpass this limitation is the integration of visual features from the speaker's lips. This is appealing since the visual channel is somewhat orthogonal to the acoustic channel, *i.e.*, the visual features are unaffected by the presence of noise environment or cross-talk among speakers.

Robust and accurate lip segmentation requires coping with several features, such as: large variation in shape and appearance across subjects caused by illumination conditions, head pose, or facial expressions. Many techniques have been proposed to achieve lip segmentation. One of the most traditional approaches rely solely on deformable models (*e.g.* [3, 4]), using both deformable model and parabolic templates [5] or deformable muscle-based face model as proposed in [6]. Improved *Active Shape Models* used in RoHiLTA approach [7] has also been proposed. The success of these bottom-up approaches are based on a number of assumptions about the imaging and motion patterns. However, the realization of these assumptions may not happen; *e.g.*, the inner/outer part of the object to be segmented may not be distinguishable as desired, or the stronger edges may not belong to the object to be segmented and tracked. Hence, given the visual non-convexity of the function being optimized in such problems, the issues above may cause the optimization procedure to get stuck in local minima. Pattern recognition methods based on regular (not deep) neural networks have been used before [8] but they still rely on the use of deformable models with threshold based processing, which is always a sensitive step, specially to salient regions that do not belong to the outer lip boundary.

In this paper we introduce a new pattern recognition based method using *Sequential Monte Carlo* (SMC) to perform lip tracking. This kind of approach holds the most competitive results in non-rigid tracking problems, but it still faces a few challenges, such as: $(i)$ the need of a large training set, $(ii)$ robustness against unseen conditions in the training set, and $(iii)$ run-time complexity of the search process. Recently, there has been a valuable effort to reduce the search complexity. For instance, in [9] the *marginal space learning* (MSL), which partitions the search space into smaller subspaces, achieves a significant complexity reduction. Another contribution [10] is a pattern recognition based approach that, given any position in the search space, the method estimates the gradient vector to optimize the segmentation function. However, the main issue with this approach is the requirement of a large training set because of the large dimensionality of the parameter space.

In this paper we address two problems of the pattern recognition methods, namely: $(i)$ robustness against imaging conditions not presented in the training set, and $(ii)$ run-time complexity of the search process. In order to handle the first issue, we propose a new observation model based on *deep belief network* (DBN) [11]. The main advantage of the DBN is its ability to provide more abstract feature spaces for classification, which can improve the robustness of the method to image conditions, and to generate optimum image features directly from the image data. In order to address the complexity issue, we resort to first and second order optimization algorithms [12]. To accomplish this, we compute the gradient and Hessian matrix directly from the output classifiers, imposing no additional requirements for the training set.

This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds and supported by project PTDC/EEA-CRO/103462/2008. *This work was partially funded by EU Project IMASEG3D (PIIF-GA-2009-236173)

## 2. PROPOSED APPROACH

Here, we aim at computing the expected contour segmentation at time instant $t$, $S_t = \{s_{i,t}\}$, where $s_{i,t} \in \Re^2$ represents the $ith$ contour sample, with $i = 1, ..., N$, given the past and current image observations, *i.e.*

$$S_t^\star = \int_{S_t} S_t \, p(S_t|I_{1:t}, y_1, \mathcal{D}) \, dS_t, \tag{1}$$

where $I_{1:t}$ denotes the set of images up to instant $t$; $\mathcal{D} = \{(I, \theta, S, K)_i\}$ is the training set containing $M$ training images $I_i$, the respective manual annotations $S_i$ and a rigid transformation $\theta = (\mathbf{x}, \gamma, \sigma) \in \Re^5$, with position $\mathbf{x} \in \Re^2$, orientation $\gamma \in [-\pi, \pi]$, and scale $\sigma \in \Re^2$; $K$ is the *lip stage*, *i.e.*, $K \in \{\text{open}, \text{semi} - \text{open}, \text{closed}\}$, (see Fig. 1 for an illustration) and $y_1$ is a random variable indicating the presence of a lip in the window defined by $\theta$ given a specific lip stage. The lip stages represent a prior information (as in [13]) that will be used in the transition model as we detail in Section 2.1.

To compute (1), we use particle filtering which approximates the filtering distribution by a weighted sum of $L$ particles and weights $\{S_t^l, w_t^{(l)}\}$, with $l = 1, \ldots, L$. Specifically, we use the *sampling importance resampling* (SIR) [14]. In the next subsections we provide details of the transition and observation models and their combination to build the proposal distribution.

**Fig. 1**. Three different snapshots taken during a speech. From left to right, the lip stages are: close, semi-open and open.

### 2.1. Transition model

From the posterior distribution in (1), we have

$$p(S_t|I_{1:t}, y_1, \mathcal{D}) \propto p(I_t|S_t, y_1, \mathcal{D}) \, p(S_t|I_{1:t-1}, y_1, \mathcal{D}), \tag{2}$$

where the transition model is

$$\begin{aligned} &p(S_t \, |I_{1:t-1}, y_1, \mathcal{D}) = \\ &\textstyle\int_{S_{t-1}} p(S_t|S_{t-1}, I_{1:t-1}, y_1, \mathcal{D}) \, p(S_{t-1}|I_{1:t-1}, y_1, \mathcal{D}) dS_{t-1}. \end{aligned} \tag{3}$$

We build the transition model as follows:

$$\begin{aligned} &p(S_t \, |S_{t-1}, I_{1:t-1}, y_1, \mathcal{D}) = \\ &\textstyle\sum_{k_{t-1}} p(S_t|S_{t-1}, K_{t-1}, y_1, \mathcal{D}) \, p(K_{t-1}|S_{t-1}, I_{t-1}, y_1, \mathcal{D}), \end{aligned} \tag{4}$$

where $p(K_{t-1}|S_{t-1}, K_{t-1}, I_{t-1}, y_1, \mathcal{D})$ is computed with the observation model (Section 2.2), and

$$\begin{aligned} &p(S_t|S_{t-1}, I_{1:t-1}, K_{t-1}, y_1, \mathcal{D}) = \\ &G(S_{t-1}|M(K_{t-1})S_{t-1}, \Sigma_S), \end{aligned} \tag{5}$$

where $M(K_{t-1})$ is a linear transformation applied to $S_{t-1}$, which is learned from the training data and $\Sigma_S$ is the covariance of the annotations also learned from the data set. In summary, the transition model is represented by a Gaussian mixture model that penalizes transitions between lip stages.

### 2.2. Observation model

The observation model from (1) is defined as:

$$p(I_t|S_t, y_1, \mathcal{D}) \propto p(S_t|I_t, y_1, \mathcal{D}) \, p(I_t|y_1, \mathcal{D}), \tag{6}$$

where the second term is assumed to be a constant and the first term is computed as follows

$$p(S_t|I_t, y_1, \mathcal{D}) = \int_\theta p(S_t|\theta, I_t, y_1, \mathcal{D}) \, p(\theta|I_t, y_1, \mathcal{D}) d\theta. \tag{7}$$

The first and the second terms in (7) are the *nonrigid* and *rigid* parts of the detection, respectively. For the computation of the nonrigid part, we assume the independence of the contour samples $s_{i,t}$, *i.e.*

$$p(S_t|\theta, I_t, y_1, \mathcal{D}) = \prod_i p(s_{i,t}|\theta, I_t, y_1, \mathcal{D}). \tag{8}$$

Defining $\psi$ as the parameter vector of the classifier for the nonrigid contour, we compute (8) as follows:

$$\begin{aligned} p(s_{i,t}|\theta, I_t, y_1, \mathcal{D}) &= \textstyle\int_\psi p(s_{i,t}|\theta, I_t, y_1, \mathcal{D}, \psi) \, p(\psi|\mathcal{D}) d\psi \\ &= \textstyle\int_\psi p(s_{i,t}|\theta, I_t, y_1, \mathcal{D}) \, \delta(\psi - \psi_{\text{MAP}}) d\psi, \end{aligned} \tag{9}$$

where $\psi_{\text{MAP}} = \arg\max_\psi p(\{S_i\}|\{(I, \theta)_i\}_{i=1..M}, \psi)$, $\delta(.)$ denotes the Dirac delta function, and $(S, I, \theta) \in \mathcal{D}$. Concerning the first probability in the result of (9), we train a regressor that indicates the most likely location of the lip border (see Fig. 2). This means that the nonrigid detection (8) can be rewritten as

$$\begin{aligned} &p(S_t|\theta, I_t, y_1, \mathcal{D}) = \\ &\textstyle\prod_i \int_\psi \delta(s_{i,t} - s_{i,t}^r(\theta, I_t, y_1, \mathcal{D})) \, \delta(\psi - \psi_{\text{MAP}}) d\psi, \end{aligned} \tag{10}$$

where, $s_{i,t}^r$ is the output of the regressor for the $i$th point. Fig. 2 shows patches used for training and testing the regressor. For instance, given an input patch like the ones displayed on the bottom right of Fig. 2, the regressor outputs the most likely location of the transition lip-skin, according to the learned model parameters $\psi_{\text{MAP}}$. Note that we also build a principal component analysis (PCA) space using the annotations $S$ from $\mathcal{D}$, and the final solution $S_t$ from (10) is obtained from a low-dimensional projection of $s_{i,t}^r$.

The rigid detection is expressed as

$$p(\theta|I_t, y_1, \mathcal{D}) \propto p(y_1|\theta, I_t, \mathcal{D}) \, p(\theta|I_t, \mathcal{D}), \tag{11}$$

where $p(\theta|I_t, \mathcal{D})$ is the prior on the space parameter. For the first term in (11) the vector of classifier parameters $\gamma$ is obtained via MAP estimation, *i.e.*, $p(\gamma|\mathcal{D}) = \delta(\gamma - \gamma_{\text{MAP}})$, so

$$p(y_1|\theta, I_t, \mathcal{D}) = \textstyle\int_\gamma p(y_1|\theta, I_t, \mathcal{D}, \gamma) \delta(\gamma - \gamma_{\text{MAP}}) d\gamma, \tag{12}$$

where $\gamma_{\text{MAP}} = \arg\max_\gamma p(y = 1|\{(I, \theta)_i\}, \gamma)_{i=1..M}$.

Note that we use DBN as the statistical model for the rigid and nonrigid classifiers described above. Fig. 2 (top) shows patches used for training the rigid classifier, Fig. 2 (middle) displays a subset of the features learned by the DBN, which resemble wavelet features, as also noticed in [11].

### 2.3. Proposal distribution

The proposal distribution is denoted as follows

$$\begin{aligned} q(S_t|S_{1:t-1}^{(l)}, I_{1:t}, y_1, \mathcal{D}) \sim \alpha q_{\text{obs}}(S_t|K_t, y_1, I_{1:t}, \mathcal{D}) \\ + (1 - \alpha) p(S_t|S_{t-1}, \mathcal{D}), \end{aligned} \tag{13}$$

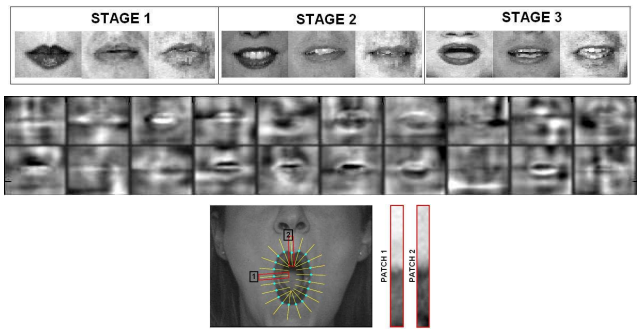


**Fig. 2**. Patches used for training the rigid classifier (top), subset of learned features (middle), normal lines from annotation points used to train the regressor (bottom).

where the first term is the observation model given by

$$q_{\text{obs}}(S_t|K_t, y_1, I_{1:t}, \mathcal{D}) = \sum_{\widetilde{S}_t} \mathcal{C}\, p(\widetilde{S}_t|I_t, y_1, \mathcal{D}) \\ G(S_t|\widetilde{S}_t, \Sigma_S), \tag{14}$$

where $\widetilde{S}_t$ denotes the set of the top detections, $\mathcal{C}$ is a normalization constant and $p(\widetilde{S}_t|I_t, y_1, \mathcal{D})$ is the probability response of the observation model of a given segmentation. The meaning of (14) is that, the higher is the overlap between the detection of the DBN and the mixture dynamical model, the larger is its weight. If there is no overlap between the DBN detection hypotheses and the mixture motion models, then the proposal distribution will be guided by the transition distribution. In this paper

$$\alpha = \max_{\widetilde{S}_t} \exp\{-K_\alpha(\widetilde{S}_t - S_{t-1})^T \Sigma_S^{-1}(\widetilde{S}_t - S_{t-1})\}, \tag{15}$$

where $K_\alpha$ is determined through cross validation.

## 3. EFFICIENT SEARCH METHODS

For the rigid and nonrigid lip segmentation (1), there is a five dimensional space and N dimensions, respectively. This results in a search space of $K^{5+N}$ samples, which is prohibitive for most practical values of $K \in [10^2, 10^3]$ and $N \in \{10, ..., 25\}$. Running the search procedure on the image pyramid, with one classifier per image scale reduces the search space substantially. The advantage here, is to reduce the number of samples in the coarser scale to $K_{\text{coarse}}$ and progressively move to finer scales only the best $K_{\text{fine}} \in [10, 30]$ candidates. Recall that, the search space procedure in fine scales needs to occur only around the current search point, which means that we have $3^5$ (3 points in 5 dimensions) samples for each point of the $K_{\text{fine}}$ positions. Furthermore, performing the nonrigid search after the rigid search, means that we have a total search space of $K_{\text{coarse}}^5 + (\#\text{scales} - 1) \times K_{\text{fine}} \times 3^5 + N$.

Our contribution is twofold: $(i)$ reduction of the search space, and $(ii)$ implementation of efficient search procedures. Concerning the first one, we learn a prior distribution from the training data on the coarse search space, and sample (via Monte Carlo sampling) $K_{\text{coarse}}$ from this distribution. This means that we have a search space with dimension $K_{\text{coarse}} + (\#\text{scales} - 1) \times K_{\text{fine}} \times 3^5 + K_{\text{fine}} \times N$. Regarding the second contribution, we propose two methods that are used in optimization algorithms: $(i)$ *gradient descent* and $(ii)$ *Newton step* [12]. This allows to reduce the exhaustive search of $3^5$ points. These two optimization methods work for convex functions, however, their use in non-convex functions, such as the ones produced by the DBN classifier, only work if the $K_{\text{coarse}}$ is sufficiently large. In gradient descent, the gradient in (12) is computed numerically using central differences, representing a computation of the classifier in 10 points of the search space, *i.e.*, $5\ \text{parameters} \times 2\ \text{points}$. Limiting the number of iterations, say, between 1 to 5 for each hypothesis, the search space is reduced to 20 to 100 points, which is smaller than $3^5$. Theoretically speaking, a faster convergence can be achieved with the Newton step, but the computation of the Hessian matrix, gradient and line search requires 25+10 search space points. Limiting the number of iterations between 1 and 5, means that the complexity of this step for one hypothesis is between 35 and 175, which is also smaller than $3^5$.

### 3.1. Training and detection procedures

For the training of the observation model, we used 1000 images (from 10 sequences) of lip annotations. For the rigid classifier, an image scale space is built, *i.e.*, $\mathcal{L}(\mathbf{x}, \boldsymbol{\sigma}) = \mathbf{G}(\mathbf{x}, \boldsymbol{\sigma}) \star \mathbf{I}(\mathbf{x})$, where $G(.)$ is the Gaussian kernel, $\mathbf{I}$ is the input image, $\boldsymbol{\sigma}$ is the image scale, and $\mathbf{x}$ is the image coordinate.

For the rigid part, we separately train three classifiers (12); one for each scale $\sigma = \{2, 4, 8\}$. The positive and negative training sets are defined based on a different scale dependent margin $m_\sigma$ that increases by a factor of two after each octave. Thus, the positive set for $\mathcal{L}(\mathbf{x}, \boldsymbol{\sigma})$, is randomly generated inside the interval $[\theta - m_\sigma/2, \theta + m_\sigma/2]$; the negative set is randomly generated outside of the interval $[\theta - m_\sigma, \theta + m_\sigma]$, where $\theta$ is the vector containing the parameters of the rigid transformation of the lips annotations. See positive patches in Fig. 2 (top).

For the nonrigid part, the regressor (10) is trained at $\sigma = 2$, where each training sample is a normal line of 41 pixels (see Fig. 2 bottom for an illustration) and the label to learn is the pixel index in the interval $[1, ..., 41]$ that is closest to the lip boundary. The structure of the DBN was determined using cross validation. Finally, we have $K_{\text{coarse}} = 10^3$ and $K_{\text{fine}} = 10$. For the SIR algorithm, the following parameters were determined through cross validation: $(i)$ *number of particles*: 100; $(ii)$ $K_\alpha = 0.1$ (see (15)).

## 4. EXPERIMENTAL RESULTS

The performance of the tracker was measured by comparing the contour estimates with the bottom-up MMDA (*Multiple Model Data Association*) tracker [15]. This tracker provides state-of-the-art results in the problem of heart tracking, which shares several of the challenges present in lip tracking (*e.g.*, varying texture and image conditions and appearance changes caused not only by motion). In MMDA, an initial contour is manually drawn in the first frame of the sequence. From this initial contour, a validation gate is built from which a discriminant Fisher classifier is trained, allowing to distinguish between lip and skin. Comparing to the MMDA, the advantages of the approach presented in this paper are the following: $(i)$ does not need an initial segmentation guess; $(ii)$ presents robustness to changing light conditions throughout the sequence and $(iii)$ does not overfit the test sequence (*i.e.*, it does not need to train a Fisher classifier for every new test image).

To evaluate the performance of the method, a manual ground truth (GT) is provided for all the images in the sequences. We use the Hammoude distance (as in [15]) to compare the contours of the manual GT and the output of the MMDA and the DBN trackers.

From the results, the proposed method exhibits advantages in the sequences containing brightness (see Fig. 3 seq. #1), low contrast between skin and lips (seq. #7), and the presence of noise (seq. #2 contains a face with a beard). However, the method still needs a representative training set. For example, seq. #3 and seq.


(a) (b) (c)

**Fig. 3**. Results obtained for the sequences #1, #2, #3, #7 and #8 with full search (a), Newton (b) and gradient (c) search methods. The quantitative results for each row is shown in Table.1.

**Table 1**. Quantitative results with the Hammoude mean distances in eight sequences.

| | GradDes | Newton | Full search | MMDA |
|---|---|---|---|---|
| $d_H$seq1 | 0.11 | **0.10** | 0.11 | 0.13 |
| $d_H$seq2 | **0.09** | 0.10 | **0.09** | 0.12 |
| $d_H$seq3 | 0.17 | 0.32 | 0.18 | **0.10** |
| $d_H$seq4 | 0.12 | 0.13 | 0.12 | **0.11** |
| $d_H$seq5 | **0.08** | **0.08** | 0.09 | 0.10 |
| $d_H$seq6 | **0.11** | 0.13 | **0.11** | **0.11** |
| $d_H$seq7 | 0.15 | 0.18 | **0.12** | 0.17 |
| $d_H$seq8 | 0.13 | 0.13 | 0.14 | **0.08** |

#8 (Fig. 3) contain sequences that are not well represented in the training set.

The run-time average complexity obtained for the search strategies in terms of floating point operations are the following: $(i)$ *Full*: $\mathcal{O}(3.5 \times 10^{11})$; $(ii)$ *GradDesc*: $\mathcal{O}(2.0 \times 10^{10})$ and $(iii)$ *Newton*: $\mathcal{O}(3.0 \times 10^{10})$.

Efficient search methods (EFM) perform as well as the full search method with a ten times smaller run-time complexity (Table 1). In fact, for sequences where the full strategy is doing well, the EFM usually improves the Hammoude distance. However, for sequences that the full method performs poorly (*e.g.*, seq. #3), the EFM worsen even more the results, which shows that robustness of our method is related to the richness of the training set.

## 5. CONCLUSIONS

In this paper we proposed a pattern recognition algorithm that can be applied to non-rigid tracking problems, such as the lip tracking. In this framework, we showed solutions to the following problems: robustness to imaging conditions and efficient search process. Regarding the robustness, the results have shown that the proposed algorithm is robust against imaging conditions noise and transition between the lips and skin. Concerning the efficiency of the search process, we have shown that the proposed search strategies allow to reduce ten times the run-time complexity while maintaining the tracking accuracy. Future work will address the reduction of the training data needed for the pattern recognition based methods.

## 6. REFERENCES


[1] R. Goecke, "Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input," in *Multimodal user interaction workshop*, vol. 57, 2005.

[2] M. Chan, Y. Zhang, and T. Huang, "Real time lip tracking and bimodal continuous speech recognition," in *IEEE* $2^{nd}$ *Wokshop Multimedia Signal Processing*, 1998, pp. 65–70.

[3] P. Aleksic, J. Williams, Z. Wu, and A. Katsaggelos, "Audiovisual speech recognition using mpeg-4 compliant visual features," *EURASIP J. Appl. Signal Processing*, vol. 11, no. 5, pp. 1213–1227, 2002.

[4] N. Paragios, "A level set approach for shape-driven segmentation and tracking of the left ventricle," *IEEE Trans. Med. Imag.*, vol. 21, no. 9, pp. 773–776, 2003.

[5] Z. Wu, P. Aleksic, and A. K. Katsaggelos, "Lip tracking for MPEG-4 facial animation," in *IEEE Int. Conf. on Multimodal Interfaces*, 2002.

[6] H. Seyedarabi, A. Aghagolzadeh, and S. Khanmohammadi, "Facial expressions animation and lip tracking using facial characteristic points and deformable model," *Int. Journal of Information Technology*, vol. 1, no. 4, pp. 416–419, 2004.

[7] L. Xie, X. Cai, Z. Fu, R. Zhao, and D. Jiang, "A robust hierarchical lip tracking approach for lipreading and audio visual speech recognition," in *Int. Conf. on Machine Learning and Cybernetics*, 2004.

[8] H. Seyedarabi, W. Lee, and A. Aghagolzadeh, "Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks," in *CCECE*, 2006, pp. 2021–2024.

[9] Y. Z. et al., "Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 392–406, 2008.

[10] S. Zhou and D. Comaniciu, "Shape regression machine," in *IPMI*, 2007, pp. 13–25.

[11] S. Salakhutdinov and G. Hinton, "Learning a non-linear embedding by preserving class neighbourhood structure," in *AI and Statisitcs*, 2007.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[13] Y.-L. Tian, T. Kanade, and J. Cohn, "Robust lip tracking by combining shape, color and motion," in *Proc. of the 4th Asian Conf. on Computer Vision*, 2000.

[14] A. Doucet, N. de Freitas, N. Gordon, and A. Smith, *Sequential Monte Carlo Methods in Practice*. Springer Verlag, 2001.

[15] J. C. Nascimento and J. S. Marques, "Robust shape tracking with multiple models in ultrasound images," *IEEE Trans. Imag. Proc.*, vol. 3, no. 17, pp. 392–406, 2008.