The Fusion of Deep Learning Architectures and Particle Filtering Applied to Lip Tracking

Gustavo Carneiro* and Jacinto C. Nascimento*
Instituto de Sistemas e Robótica
Instituto Superior Técnico, Lisbon, Portugal

Abstract

This work introduces a new pattern recognition model for segmenting and tracking lip contours in video sequences. We formulate the problem as a general non-rigid object tracking method, where the computation of the expected segmentation is based on a filtering distribution. This is a difficult task because one has to compute the expected value using the whole parameter space of segmentation. As a result, we compute the expected segmentation using sequential Monte Carlo sampling methods, where the filtering distribution is approximated with a proposal distribution to be used for sampling. The key contribution of this paper is the formulation of this proposal distribution using a new observation model based on deep belief networks and a new transition model. The efficacy of the model is demonstrated in publicly available databases of video sequences of people talking and singing. Our method produces results comparable to state-of-the-art models, but showing potential to be more robust to imaging conditions.

1 Introduction and Prior work

The automatic lip segmentation is a challenging problem, which is often used as a testbed to demonstrate the efficacy of systems designed to track highly deformable structures. The problem is difficult due to several challenges, for example, high variability of the shapes, colors, textures, and changing lighting conditions. Lip tracking is important in audio-visual speech recognition systems, such as the *automatic speech recognition* (ASR), which has been widely deployed in mobile phones and car environments. Most ASR systems have concentrated exclusively on the acoustic speech signal, which means that they are susceptible to acoustic noise. Incorporating automatic lip reading from visual information can improve the performance of these systems, namely, when the acoustic signal is corrupted. This technique can also exploit additional information contained in lip motion. Compared to conventional acoustic recognition, audio-visual speech recognition systems can decrease the *Word Error Rate* (WER) for various signal/noise conditions [10, 7].

Different techniques have been proposed for lip tracking. However, the majority of the methods are not suited for an accurate lip segmentation when the image conditions (*e.g.* illumination, texture) changes over time. For instance, deformable models [1] use a prior information about the foreground and background, which is potentially unreliable since the image conditions can vary over and between sequences. Optical flow provides good performance in tracking tasks [4], but its constraint can be violated in situations where changes are caused by the appearance (and not motion). Methods based on active shape models (ASM) [2] do not provide the best results in situations containing non-rigid and large textures changes. The Flexible Eigentracking [3] extends ASM to handle large non-rigid lip deformations, but the lack of more powerful dynamical models may represent a problem. Tracking non-rigid visual objects using particle filtering has produced excellent results [9, 13], but the main focus has been on formulating low-dimensional state spaces, which is an important problem, but orthogonal to what is proposed in this paper.

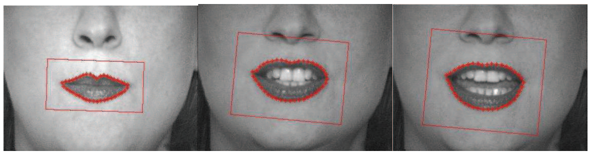

Figure 1. Three different snapshots taken during a speech with the rigid (rectangle) and non-rigid (lip contour) annotations. From left to right, the lip stages are: close, semi-open and open.

In this paper, we introduce a new pattern recognition model using *Sequential Monte Carlo* (SMC) methods to track highly deformable visual objects. Specifically, we apply our method to the problem of lip tracking. Regarding this model our main contributions are: (i) a new transition model, (ii) a new observation model based on deep belief networks (DBN), and (iii) a formulation of a new proposal distribution. The transition model pro-

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

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.508

posed in this paper makes use of the prior information about the lip stage, as in [12]. Specifically, we consider three different lip stages: close, semi-open or open (Fig. 1). Also, the deformation caused by these motion patterns are described by a linear transform, whose parameters are learned from the training data. Concerning the observation model, this is based on a DBN architecture, which involves a statistical pattern recognition model [11]. The main advantage of deep belief networks is its ability to produce more abstract feature spaces for classification which has the potential to improve the robustness of the method to image conditions and to generate optimum image features (in terms of classification) directly from image data. Finally, our proposal distribution, inspired by the work in [8], combines the detection results from the deep learning architecture with the transition model. The system based on this model produces precise lip segmentation and tracking, and shows robustness to imaging conditions. We show quantitative comparisons between our method and a non-rigid state-of-the-art tracking approach [6] on publicly available data sets.

2 Proposed approach

Our main goal is to compute the expected segmentation at time instant t, $S_t = \{s_{i,t}\}_{i=1..N}$, where $s_{i,t} \in \Re^2$ represents the segmentation points, *i.e.*

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

where $I_{1:t}$ denotes the set of images up to instant t; $\mathcal{D} = \{(I, \theta, S, K)_i\}_{i=1..M}$ is the training set containing M training images I_i, the respective manual annotations S_i and rigid transformations $\theta_i = (\mathbf{x}, \gamma, \sigma) \in \Re^5$, with position $\mathbf{x} \in \Re^2$, orientation $\gamma \in [-\pi, \pi]$, and scale $\sigma \in \Re^2$ (Fig. 1 displays the window representing the rigid transform); y_1 is a random variable indicating the presence of a lip using a rigid transformation θ, and K is the *lip stage*, *i.e.*, $K \in \{\mathrm{open}, \mathrm{semi-open}, \mathrm{closed}\}$, as shown in Fig. 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) [5]. In the next subsections we provide details of the transition and observation models and their combination to build the proposal distribution.

2.1 Transition model

From the posterior distribution in (1), we have

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

where the prediction model is defined as

$$p(S_t | I_{1:t-1}, y_1, \mathcal{D}) = \qquad (3)$$
$$\int_{S_{t-1}} p(S_t | S_{t-1}, I_{1:t-1}, y_1, \mathcal{D}) \, p(S_{t-1} | I_{1:t-1}, y_1, \mathcal{D}) dS_{t-1}.$$

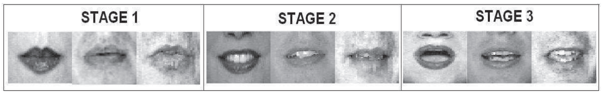

Figure 2. Training images for each lip stage.

Figure 3. Subset of learned features.

We build the transition model as follows:

$$p(S_t | S_{t-1}, I_{1:t-1}, y_1, \mathcal{D}) = \qquad (4)$$
$$\sum_{K_{t-1}} p(S_t | S_{t-1}, K_{t-1}, I_{1:t-1}, y_1, \mathcal{D}) \, p(K_{t-1} | S_{t-1}, I_{t-1}, y_1, \mathcal{D})$$

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

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

where $M(K_{t-1})$ is a linear transformation applied to S_{t-1}, which is learned from the training data and Σ_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 (2) is defined as:

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

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

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

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.*

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

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

$$p(s_{i,t} | \theta, I_t, y_1, \mathcal{D}) = \int_\psi p(s_{i,t} | \theta, I_t, y_1, \mathcal{D}, \psi) \, p(\psi | \mathcal{D}) d\psi \qquad (9)$$
$$= \int_\psi p(s_{i,t} | \theta, I_t, y_1, \mathcal{D}) \, \delta(\psi - \psi_{MAP}) d\psi$$

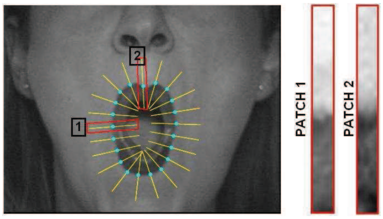

Figure 4. Lines perpendicular to annotation points used to form the patches (patches 1 and 2 are shown on the right) to train the regressor.

where $\psi_{MAP} = \arg\max_\psi p(\{S_i\}|\{(I,\theta)_i\}_{i=1..M}, \psi)$, $\delta(.)$ 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. 4). This means that the nonrigid detection (8) can be in practice rewritten as

$$p(S_t|\theta, I_t, y_1, \mathcal{D}) = \prod_i \int_\psi \delta(s_{i,t} - s^r_{i,t}(\theta, I_t, y_1, \mathcal{D}))\, \delta(\psi - \psi_{MAP}) d\psi \tag{10}$$

where, $s^r_{i,t}$ is the output of the regressor for the ith point. Fig. 4 shows patches used for training and testing the regressor. For instance, given an input patch like the ones displayed on the right of Fig. 4, the regressor outputs the most likely location of the transition lip-skin, according to the learned model parameters ψ_{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 (11) is obtained from a low-dimensional projection of $s^r_{i,t}$.

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 γ is obtained via MAP estimation, *i.e.*, $p(\gamma|\mathcal{D}) = \delta(\gamma - \gamma_{MAP})$, so

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

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

Note that we use DBN as the statistical model for the rigid and nonrigid classifiers described above. Fig. 2 shows patches used for training the rigid classifier and Fig. 3 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

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

the first term in (13) 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; $p(\widetilde{S}_t|I_t, y_1, \mathcal{D})$ is the probability response of the observation model of a given segmentation. The meaning of (13) is that, the higher is the overlap between the detection of the DBN and the mixture dynamical model, the larger is its weight on the proposal. 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_α is determined through cross validation.

3 Experimental Results

In this section we provide a comparison between the proposed method and the MMDA (*Multiple Model Data Association*) tracker [6]. 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 (orthogonal lines radiating from these points) is built from which a discriminant Fisher classifier is trained [1], 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 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 [6]) to compare the contours of the manual GT and the output of the trackers. The distance is defined as follows

$$d_H(\mathcal{X}, \mathcal{S}) = \frac{\#((R_\mathcal{X} \cup R_\mathcal{S}) - (R_\mathcal{X} \cap R_\mathcal{S}))}{\#(R_\mathcal{X} \cup R_\mathcal{S})} \tag{16}$$

where $R_\mathcal{X}$ represents the image region by the contour $\mathcal{X}$, and similarly for $R_\mathcal{S}$.

Table 1 shows the performance in terms of the Hammoude distance in eight test sequences, each containing around 100 images showing large appearance changes, rigid and non-rigid deformations. The training set consists of 10 sequences, each containing around 100 images. Fig. 5 shows the results of the tracking method in four images of each of the eight test sequences.

Table 1. Mean Hammoude distance of our method and MMDA [6] in 8 sequences.

	Our Method	MMDA[6]
$d_H$seq1	**0.11**	0.13
$d_H$seq2	**0.09**	0.12
$d_H$seq3	0.18	**0.10**
$d_H$seq4	0.12	**0.11**
$d_H$seq5	**0.09**	0.10
$d_H$seq6	**0.11**	**0.11**
$d_H$seq7	**0.12**	0.17
$d_H$seq8	0.14	**0.08**

From the results, we see that in general, MMDA [6] works well in sequences that show a well-defined transition between lip and skin, which is the case of all sequences, except for Seq. 7, where our method presents much better results, demonstrating a larger robustness to image conditions. Another advantage of our approach is shown in Seq. 2, where a face appearance containing a beard represents an issue for MMDA, but not for our method. However, in cases where that training set does not cover the variations present in the test set, then our method does not work so well (see Seq. 3 and Seq. 8).

4 Conclusions

In this paper we propose a new tracking algorithm that can be applied to non-rigid tracking problems, such as the lip tracking. Using a Sequential Monte Carlo sampling algorithm, our main contributions are a new transition and observation models, and a new proposal distribution. The experiments show competitive tracking results, which are compared quantitatively to a state-of-the-art tracking approach. The combination of different types of models in the proposal distribution and the use of deep belief networks provide accuracy and robustness to imaging conditions and drifting.

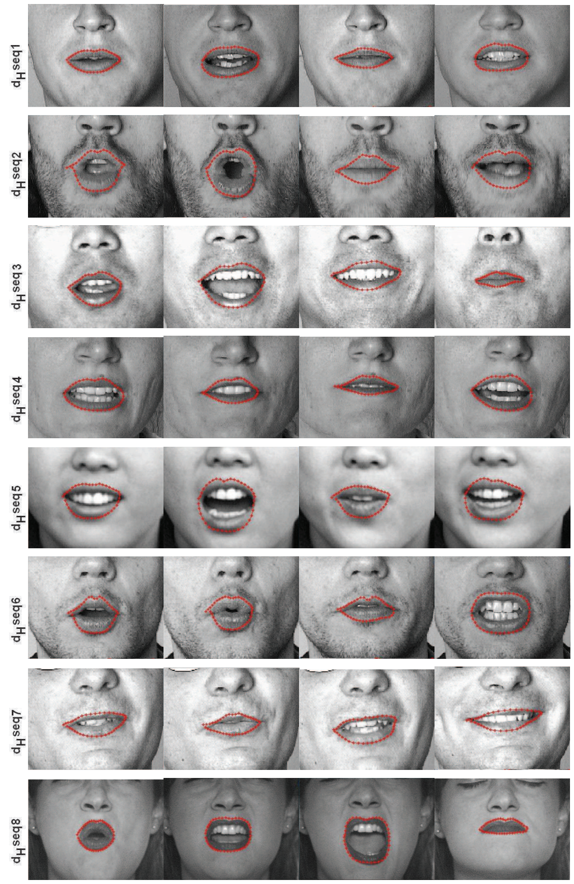

Figure 5. Results obtained with the tracking method introduced in this paper. The quantitative result for each row is shown in Tab.1.

References

[1] A. Blake and M. Isard. *Active Contours*. Springer-verlag, Oxford, 1998.

[2] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Eur. Conf. Comp. Vis.*, pages 484–498, 1998.

[3] F. de la Torre, J. Melenchon, J. Vitria, and P. Radeva. Eigenfiltering for flexible eigentracking (efe). *Int. Conf. Pattern Recognition*, 3:7118, 2000.

[4] D. DeCarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Comp. Vis.*, 38(2):99–127, 2000.

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

[6] J. C. Nascimento and J. S. Marques. Robust shape tracking with multiple models in ultrasound images. *IEEE Trans. Imag. Proc.*, 3(17):392–406, 2008.

[7] A. Nefian, L. H. Liang, L. Xiao, X. X. Liu, X. Pi, and K. Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal of Applied Signal Processing*, (11):1274–1288, December 2002.

[8] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Eur. Conf. Comp. Vis.*, 2004.

[9] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *IEEE Int'l Conf. Face and Gesture Recognition*, 2004.

[10] G. Potamianos, C. Neti, G. Gravier, and A. Garg. Recent advances in the automatic recognition of audio-visual speech. *Proceedings of the IEEE*, 91(9), 2003.

[11] R. Salakhutdinov and G. Hinton. Learning a non-linear embedding by preserving class neighbourhood structure. *AI and Statistics*, 2007.

[12] Y. Tian, T. Kanade, and J. Cohn. Robust lip tracking by combining shape, color and motion. In *Proc. of the Asian Conf. on Computer Vision*, 2000.

[13] Q. Wang, G. Xu, and H. Ai. Learning object intrinsic structure for robust visual tracking. In *Conf. Vis. Patt. Rec.*, 2003.