TEMPORAL MODELING OF FACIAL ACTIONS FROM FACE PROFILE IMAGE SEQUENCES

*Maja Pantic and Ioannis Patras**
Delft University of Technology
EEMCS / Mediamatics Dept.
Delft, the Netherlands
{M.Pantic,I.Patras}@ewi.tudelft.nl

ABSTRACT

The recognition of facial action units (AUs) in image sequences is a challenging problem. AU detectors achieve good recognition rates, but virtually all of them deal only with frontal-view face images and cannot handle temporal dynamics of AUs. In this work we report on a system for automatic recognition of temporal models of AUs from long, profile-view face image sequences. We exploit particle filtering to track 15 facial points in an input face-profile video sequence and we introduce facial-behavior temporal-dynamics recognition from continuous video input using temporal rules. The utilized algorithm performs both automatic segmentation and recognition of temporal segments (i.e., onset, apex, offset) of 23 AUs occurring alone or in a combination in an input face-profile video sequence. A recognition rate of 88% is achieved.

1. INTRODUCTION

Facial expression is one of the most cogent, naturally preeminent means for human beings to communicate emotions, to clarify and stress what is said, to signal comprehension, disagreement, and intentions, in brief, to regulate interactions with the environment and other persons in the vicinity [1, 2]. Automatic analysis of facial expression attracted, therefore, the interest of many AI researchers – automated systems will have numerous applications in behavioral science, medicine, security, and human-computer interaction.

Most approaches to automatic facial expression analysis attempt to recognize a small set of prototypic emotional facial expressions, such as sad, angry, surprised and happy [3]. Yet such prototypic expressions occur relatively infrequently. Typically displayed facial expressions often convey signs of attitudinal states such as interest and boredom, conversational signals, and blends of two or more affective states [1]. Instead of classifying facial expressions into few basic emotion categories, this work attempts to measure a large range of facial behavior by recognizing facial actions (i.e., atomic facial signals) that produce expressions.

The method proposed here is based on the Facial Action Coding System (FACS) [4]. It is a system designed for human observers to describe changes in facial expression in terms of observable facial muscle actions (i.e., facial action units, AUs). FACS provides the rules for visual detection of 44 different AUs and their temporal segments (onset, apex, offset) in a video of an observed face. Using these rules, a human coder decomposes a shown facial expression into the specific AUs that produced the expression. Hence, AUs can be seen as being analogous to phonemes for facial expression.

Few methods were reported for automatic AU detection in face image sequences and none was reported for automatic recognition of temporal dynamics of AUs [5]. Although FACS is the leading method for measuring facial behavior in behavioral science, achieving AU recognition by computer remains difficult. A problem is that AUs can occur in more than 7000 combinations, causing various in- and out-of-image-plane movements of facial components (e.g., pursed lips, jaw dropped, jetted jaw) that are difficult to detect from a single 2D facial-view. The analysis of multiple views of the face has been identified as a promising approach to solving both this and the problem of pose variability that the inevitable presence of head movements imposes [6, 3]. Nevertheless, most of the existing AU detectors deal only with frontal-view face image sequences. For example, Cohn et al. [7] presented a method based on facial feature point tracking that can recognize 8 individual AUs and 7 combinations of AUs in frontal-view face image sequences free of head motions. Tian et al. [8] presented a system based upon lip tracking and template matching that recognizes 16 AUs occurring alone or in a combination in a nearly frontal-view video of the face. Bartlett et al. [9] reported on automatic recognition of 3 AUs using Gabor filters, support vector machines, and hidden Markov models to analyze an input nearly frontal-view face image sequence.

In contrast to this past work on automatic AU detection, which deals only with frontal-view face images and cannot recognize temporal dynamics of AUs, we introduce here automatic detection of AUs and their temporal dynamics from profile-view face image sequences. We carried out this research with three motivations:

1. In a frontal view of the face, AUs such as puckering the lips (AU18) or pushing the jaw forwards (AU29) represent out-of-image-plane non-rigid facial movements which are difficult to detect [8]. Such AUs are clearly observable in a profile-view of the face [10].
2. Temporal dynamics of AUs (i.e., the timing and the duration of facial activity) is a critical factor for the interpretation of the observed facial behavior [1]. Nevertheless, no effort towards automating the detection of the temporal segments of AUs in face image sequences has been reported so far.
3. A basic insight in how automatic AU detection from a profile-view of the face can be achieved is necessary if a technological framework for automatic AU detection from multiple views of the face is to be established.

Fig. 1 outlines our method, a preliminary version of which was reported in [10]. This previous version had several limitations: it did not use temporal cues, it did not handle recognition of temporal dynamics of AUs, and AU coding was based only upon changes in the contour of the face profile region (i.e., changes within the face profile region were disregarded). The current version of the method addresses these limitations. It operates under two assumptions: (1) the input video sequence is non-occluded nearly left profile view of the face with possible in-image-plane head rotations, and (2) the first frame of it shows a neutral expression and no head rotations. After the fiducial points are initialized in the first frame of the input face profile image sequence, we exploit particle filtering to track

* The work of M. Pantic and I. Patras is supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202.

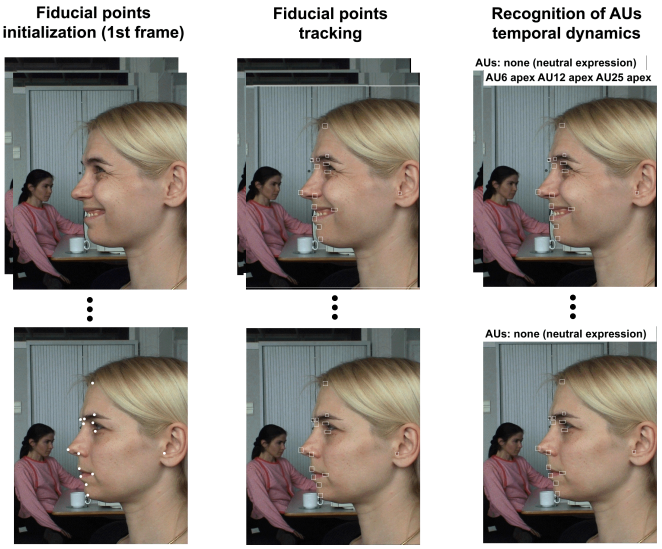

Fig. 1: Recognition of AU temporal dynamics from profile-view face image sequences

these 15 points automatically for the rest of the sequence. Based upon the changes in the position of the fiducial points, we measure changes in facial expression. Changes in the position of the fiducial points are transformed first into a set of mid-level parameters for AU recognition. Based upon the temporal consistency of mid-level parameters, a rule-based method encodes temporal segments (onset, apex, offset) of 23 AUs occurring alone or in a combination in input nearly-left-profile-views of the face. The usage of temporal information allows us not only to code a video segment to the corresponding AUs, but also to automatically segment an arbitrary long video sequence to the segments that correspond to different expressions. Fiducial-point tracking, parametric representation, AU coding, automatic segmentation of the video sequence, and experimental evaluation are explained in sections 2, 3, 4, 5 and 6.

2. FIDUCIAL-POINT TRACKING

Facial muscle activity produces changes in the appearance of the facial features (eyes, nose, lips, etc.); their shape and location can alter immensely with facial expressions (e.g., pursed lips vs. jaw dropped). To reason about the shown facial expression and about the facial muscle actions that produced it, we track a set of 15 facial fiducial points (Fig. 2), the location of which alters during the facial expressions. At the first frame of the sequence, a number of windows that are interactively positioned around each of the facial fiducial points, define a number of color templates. Let us denote such a color template with $\boldsymbol{o} = \{\boldsymbol{o_i}\}$ where $\boldsymbol{i}$ is the pixel subscript. We subsequently track each color template for the rest of the image sequence with the auxiliary particle filter that was introduced by Pitt and Shepard [11]. Particle filtering has become the dominant tracking paradigm due to its ability to deal successfully with noise, occlusion and clutter. In order to adapt it for the problem of color-based template tracking, we define an observation model that is based on a robust color-based distance between the color template $\boldsymbol{o} = \{\boldsymbol{o_i} \mid \boldsymbol{i} = 1\ldots M\}$ and a color template $\boldsymbol{c} = \{\boldsymbol{c_i} \mid \boldsymbol{i} = 1\ldots M\}$ at the current frame. We attempt to deal with shadows by compensating for the global intensity changes and with outliers by using robust error norms. The latter is particularly important because, when tracking profile facial points, a part of the template is bound to contain information from the background. We use the distance function d given in *(1),* where M is the number of pixels in each template, $\boldsymbol{m_c}$ (and $\boldsymbol{m_o}$) is the average intensity of template $\boldsymbol{c} = \{\boldsymbol{c_i}\}$ (and, respectively, of template $\boldsymbol{o} = \{\boldsymbol{o_i}\}$), $\boldsymbol{i}$ is the pixel index and the

$$d = \sum_{i=1}^{M} \boldsymbol{\rho}\left(\left\|\frac{c_i}{m_c} - \frac{o_i}{m_o}\right\|_1, \boldsymbol{\mu_c}\right) / M \rightarrow (1)$$

robust function that we use is the absolute value.

We proceed under 2 assumptions: (1) the input image sequence is non-occluded nearly left profile view of the face with possible in-image-plane head rotations, and (2) the first frame shows a neutral expression and no head rotations. To handle possible in-image-plane head rotations and variations in scale of the observed face profile, we register each frame of the input image sequence with the first frame based on two referential points (Fig. 2): the tip of the nose (P4) and the top of the forehead (P1). We use these points as the referential points because of their stability with respect to non-rigid facial movements: facial muscle contractions do not cause physical displacements of these points. The current frame $\boldsymbol{t}$ is registered with the first frame $\boldsymbol{t1}$ so that the line P1P4 discerned for frame $\boldsymbol{t}$ is of the same length and orientation as the line P1P4 determined for the first frame $\boldsymbol{t1}$. Except of P1 and P4, other facial fiducial points are tracked in the registered input image sequence. Typical results are illustrated in Fig. 1.

3. MID-LEVEL PARAMETRIC REPRESENTATION

Contractions of facial muscles alter the shape and location of the facial features. Some of these changes in facial expression are observable from the changes in the position of the tracked points. To classify the tracked changes in terms of AUs, these changes are transformed first into a set of mid-level parameters.

We defined three mid-level parameters in total: ***up/down(P)***, ***in/out(P)***, and ***inc/dec(PP')***. Parameter ***up/down(P)*** = $y(P_{t1}) - y(P_t)$ describes upward and downward movements of point P. If $y(P_{t1}) - y(P_t) > \varepsilon$, point P moves up. If $y(P_{t1}) - y(P_t) < \varepsilon$, point P moves down. P_{t1} is point P localized in the first frame of the input image sequence. P_t is point P tracked in frame t. The value of $y(P)$ is the y-coordinate of point P and the value of ε is 1 pixel. Parameter ***in/out(P)*** = $x(P_{t1}) - x(P_t)$ describes inward and outward movements of point P. If $x(P_{t1}) - x(P_t) < \varepsilon$, point P moves inward. If $x(P_{t1}) - x(P_t) > \varepsilon$, point P moves outward. Parameter ***inc/dec(PP')*** = $PP'_{t1} - PP'_t$ describes the increase or decrease of the distance between points P and P'. If $PP'_{t1} - PP'_t < \varepsilon$, distance PP' increases. If PP'_{t1}

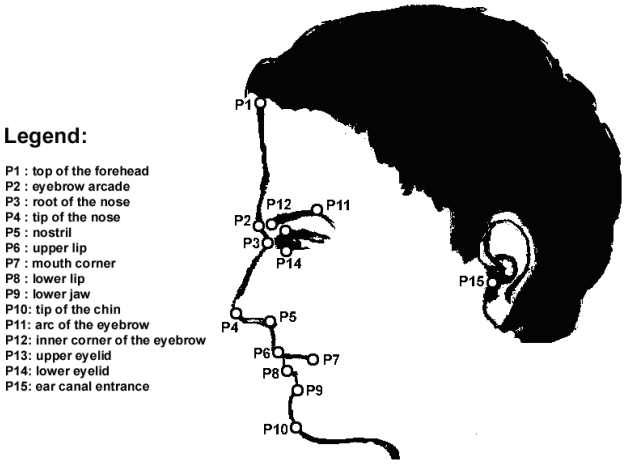

Fig. 2: Facial fiducial points

Table 1: Mid-level parameters for AU recognition

	Parameters		*Parameters*
AU1	up/down(P2) > ε	AU2	up/down(P11) > ε
AU4	inc/dec(P2P12) > ε	AU5	inc/dec(P13P14) < ε
AU6	t1 >inc/dec(P13P14) > ε up/down(P7) > ε	AU7	t1 >inc/dec(P13P14) > ε |up/down(P7)| ≤ ε
AU9	inc/dec(P2P3) > ε	AU10	inc/dec(P5P6) > ε in/out(P6) > ε
AU12	up/down(P7) > ε inc/dec(P7P15) > ε	AU13	up/down(P7) > ε |inc/dec(P7P15)| ≤ ε
AU15	up/down(P7) < ε	AU17	in/out(P10) < ε
AU16	up/down(P8) < ε inc/dec(P8P10) > ε	AU18	inc/dec(P7P15) < ε in/out(P8) > ε
AU20	inc/dec(P7P15) > ε	AU23	t2 > inc/dec(P6P8) > ε
AU24	inc/dec(P6P8) > t2	AU25	inc/dec(P6P8) < ε |inc/dec(P4P10)| ≤ ε
AU26	t3 > inc/dec(P4P10) > ε	AU27	inc/dec(P4P10) > t3
AU29	in/out(P10) > ε	AU36[b]	in/out(P9) > ε
AU44	inc/dec(P13P14) > t1		

– PP'_t > ε, distance *PP'* decreases. Distance *PP'* is calculated as the Euclidian distance between points *P* and *P'*. These mid-level parameters are calculated for various points, for each input frame.

4. ACTION UNIT RECOGNITION

To code an input face-profile image sequence in terms of 23 AUs, occurring alone or in a combination, we use a dynamic approach that employs temporal information to discriminate different AUs. The logic behind using the temporal information is that each AU has a unique temporal pattern. To minimize the effects of noise and inaccuracies in fiducial point tracking and to enable the recognition of temporal patterns of shown AUs, the utilized approach considers the temporal consistency of the mid-level parameters.

We divide activation of each AU into three temporal segments: the onset (beginning), apex, and offset (ending). Each temporal rule utilized for AU recognition is further defined in terms of the mid-level parameters (for the full list of mid-level parameters utilized to discriminate 23 different AUs, see Table 1) and each encodes a specific temporal segment of a single AU in a unique way. For example, to recognize the temporal segments of AU1, which causes upward movement of the inner corners of the eyebrows, we exploit the following temporal rules (ε is 1 pixel):

IF ($[up/down(P2)]_t > [up/down(P2)]_{t-1} + \varepsilon$)
AND $up/down(P2) > \varepsilon$ THEN ***AU1-onset***

IF $|[up/down(P2)]_t - [up/down(P2)]_{t-1}| \leq \varepsilon$
AND $up/down(P2) > \varepsilon$ THEN ***AU1-apex***

IF ($[up/down(P2)]_t < [up/down(P2)]_{t-1} - \varepsilon$)
AND $up/down(P2) > \varepsilon$ THEN ***AU1-offset***

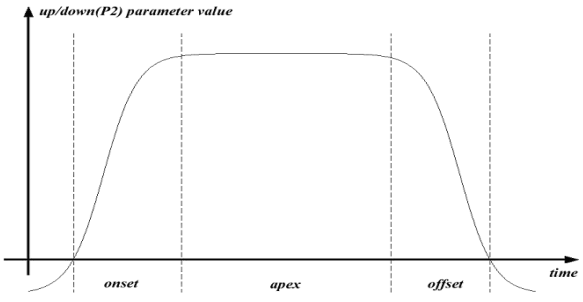

Fig. 3: The temporal pattern of AU1 activation

Fig. 3 illustrates the meaning of these rules. The horizontal axis represents the time dimension (i.e., the image sequence) and the vertical axis represents values that the parameter *up/down*(P2) can take. Since the upward motion of the skin surface of the eyebrow arcade is the principle cue for the activation of AU1, the upward movement of the fiducial point P2 (i.e., *up/down*(P2)) is used as the criterion for detecting the onset of the AU1 activation. Reversal of this motion parameter is used to detect the offset of this facial expression. Since Fig. 3 represents an abstraction to the typical progression of AU1, specific *up/down*(P2) parameter values are not provided. Fig. 3 indicates that P2 should be moving upward and it should be above its neutral-expression location to label a frame with an "AU1 onset". The upward motion should terminate, resulting in a stable temporal location of P2, before a frame can be labeled as "AU1 apex". Eventually, P2 should move downward toward its neutral-expression location to label a frame as an "AU1 offset". Generally, for each and every AU, it must be possible to detect a temporal segment (an onset, apex, or offset) continuously over at least 5 consecutive frames for the facial action in question to be scored. Incited by the research findings that suggested that temporal changes in neuromuscular facial activity last from ¼ of a second (e.g., a blink) to several minutes (e.g., a jaw clench) [4], the utilized temporal duration has been determined empirically based on a video frame rate of 24 frames/second (i.e., 5 frames have a duration of less than ¼ of a second).

Both inaccuracies in facial point tracking and occurrences of non-prototypic facial activity may result in temporal segments that are unlabeled (i.e., neither the onset, nor the apex, nor the offset) or in frames and temporal segments that are labeled incorrectly. The latter may arise, for instance, when an apex frame or an apex temporal segment of an AU is detected either between two onset segments or between two offset segments of that AU. To handle such situations, we employ a memory-based process that takes into account the dynamics of facial expressions. More specifically, we examine the labels of both the previous and next frame / segment and re-label the current frame / segment according to the ruled-based system summarized in Table 2. For instance, any unlabeled temporal segment and/or any apex segment of an AU that has been detected between two onset segments of that AU are re-labeled as "onset". Finally, an AU should be recognized, in general, only when the full temporal model of that AU is observed (e.g., see Fig. 3 for the case of AU1). Yet, in order to deal with fast transitions between onset and offset temporal segments, we score AUs even if the relevant apexes are missing.

5. AUTOMATIC SEGMENTATION

Virtually all the existing AU detectors perform well only on isolated or pre-segmented facial expression image sequences (i.e., picturing a

Table 2: Rules for resolving temporal conflicts/uncertainties.
R3 is not used if a single frame is unlabeled. It is only used if a temporal segment (a sequence of at least 5 consecutive frames) of an AU is unlabeled. The rest of rules are used for both frames and temporal segments that are unlabeled or labeled incorrectly.

	Previous labeling	*Current (old label)*	*Subsequent labeling*	*Current (new label)*
R1	Onset	Unlabeled / Apex	Onset	Onset
R2	Onset	Unlabeled	Apex	Apex
R3	Onset	Unlabeled	Offset	Apex
R4	Apex	Unlabeled	Apex	Apex
R5	Apex	Unlabeled	Offset	Apex
R6	Offset	Unlabeled / Apex	Offset	Offset

Table 3: AU recognition results. Upper face AUs: AU1, AU2, AU4, AU5-AU7, AU9, AU44. AUs affecting the mouth: AU10, AU12, AU13, AU15, AU16, AU18, AU20, AU23-AU25. AUs affecting the jaw: AU17, AU26, AU27, AU29, AU36. # denotes the number of samples. C denotes correctly recognized samples. MA denotes the number of samples in which some AUs were missed or they were scored in addition to those depicted by human experts. IC denotes incorrectly recognized samples.

	#	C	MA	IC	Rate
upper face	61	55	6	0	90.1%
mouth	44	39	3	2	88.6%
jaw	23	21	2	0	91.3%
all 23 AUs	68	60	6	2	**88.2%**

single temporal activation pattern of either a single AU or an AU combination). In reality, such segmentation is not available and, hence, there is a need to find an automatic way of segmenting face image sequences into the different facial expressions pictured.

To automatically segment an arbitrary long video sequence to the segments that correspond to different facial expressions, we use a sequential facial expression modeling that employs information on shown temporal patterns of AUs. The display of a certain expression in video corresponds to a temporal sequence of facial motions that we represent as a sequence of temporal patterns (onset-apex-offset) of one or more AUs. It seems natural to model this sequential event with a model that also starts from a fixed starting occurrence, always reaches an end occurrence, and has the probability of changing the occurrence sequence set to zero. Since the presence of facial activity determines the shown facial expression, its absence can be used to delimit the transition between different facial expressions. The term "neutral facial expression" is usually used to designate the absence of facial activity. So, to solve the segmentation problem, we use a neutral-expressive-neutral sequential facial expression model, where "expressive" segment contains temporal patterns (onset-apex-offset) of one or more AUs encoded by our AU recognition method.

6. EXPERIMENTAL EVALUATION

Though AU-coded facial expression image databases are available in general, these databases contain portraits or nearly frontal-views of human faces. Since these data are not suitable for testing our face-profile-based AU encoder, we generated our own test data.

The test data set has been created in office environments (Fig. 1). It includes 34 face-profile image sequences of 6 different faces of subjects of both sexes (50% female), ranging in age (20 to 42 years), and ethnicity. The subjects were instructed to display series of expressions (2-5 expressions; 160-540 frames), each of which included a single AU or an AU combination as well as a neutral state at the beginning and at the end of it. The size of the face region in each frame was at least 300 pixels across the width of the face. Sequences began with a neutral state with no head rotation. 2 expert FACS coders were asked to depict displayed AUs and their temporal segments in each of the 82 facial expressions constituting 34 face-profile image sequences of our data set. They agreed on the AUs displayed in 68 facial expressions. The AU-coded descriptions of these 68 expressions given by the two human FACS coders were compared further to those produced by our method for the automatically segmented input image sequences. The results of this comparison are given in Table 3. Most of the misidentifications produced by our method arose from confusion between similar AUs (e.g., AU6 and AU7, AU12 and AU13, AU25 and AU26). As far as the recognition of temporal segments of AUs is concerned, the temporal segments indicated by 2 human experts were delayed for few frames in comparison to those detected by our method.

7. CONCLUSIONS

In this paper we presented a new method for automatic recognition of temporal models of AUs from long, profile-view face image sequences. The proposed method extends the state of the art in the field in several directions, including the facial view (profile), the segmentation of an arbitrary long video sequence to the segments corresponding to different expressions, the temporal modeling of AU activation, and the number of AUs (i.e., 23) handled. Namely, the previously reported AU detectors do not deal with the profile view of the face, cannot handle continuous facial expression image sequences, cannot manage temporal dynamics of AUs, cannot detect out-of-image-plane non-rigid facial movements such as pushing the jaw forward (AU29) and, at the best, can recognize 16 AUs (from in total 44 AUs) occurring alone or in a combination in a face image sequence [8].

Nonetheless, the proposed algorithm has some limitations. For example, it cannot analyze face profile image sequences of subjects having facial hair (due to occlusion of feature points P5-P10, P15) or loose hair on the forehead (due to occlusion of referential point P1). Also, though the proposed method demonstrates an acceptable level of concurrent validity with manual FACS coding of test data (Table 3), additional field trials (i.e., a larger set of test images with more subjects) and more elaborate quantitative validation studies are necessary to confirm this finding.

REFERENCES

[1] J. Russell and J. Fernandez-Dols, *The psychology of facial expression*, Cambridge University Press, 1997.

[2] D. Keltner and P. Ekman, "Facial expression of emotion", *Handbook of Emotions*, Guilford Press, pp. 236-249, 2000.

[3] M. Pantic and L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal HCI", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.

[4] P. Ekman and W. Friesen, *Facial Action Coding System*, Consulting Psychologist Press, 1978.

[5] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art", *IEEE TPAMI*, vol. 22, no. 12, pp. 1424-1445, 2000.

[6] A. Pentland, "Looking at people", *IEEE TPAMI*, vol. 22, no. 1, pp. 107-119, 2000.

[7] J.F. Cohn, et al., "Automated face analysis by feature tracking has high concurrent validity with manual face coding", *Psychophysiology*, vol. 36, pp. 35-43, 1999.

[8] Y. Tian, et al., "Recognizing action units for facial expression analysis", *IEEE TPAMI*, vol. 23, no. 2, pp. 97-115, 2001.

[9] M.S. Bartlett, et al., "A prototype for automatic recognition of spontaneous facial actions", *Advances in Neural Information Processing Systems*, vol. 15, 2003.

[10] M. Pantic, et al., "Facial action recognition in face profile image sequences", *IEEE ICME*, pp. 37-40, 2002.

[11] M.K. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filtering", *J. Amer. Stat. Assoc.*, vol. 94, pp. 590-599, 1999.