J Braz Comput Soc (2013) 19:3–13
DOI 10.1007/s13173-012-0081-7

ORIGINAL PAPER

A model for inference of emotional state based on facial expressions

Rafael A. M. Gonçalves · Diego R. Cueva · Marcos R. Pereira-Barretto · Fabio G. Cozman

Received: 24 October 2011 / Accepted: 19 May 2012 / Published online: 24 June 2012

Abstract Non-verbal communication is of paramount importance in person-to-person interaction, as emotions are an integral part of human beings. A sociable robot should therefore display similar abilities as a way to interact seamlessly with the user. This work proposes a model for inference of conveyed emotion in real situations where a human is talking. It is based on the analysis of instantaneous emotion by Kalman filtering and the continuous movement of the emotional state over an Emotional Surface, resulting in evaluations similar to humans in conducted tests. A simulation-optimization heuristic for system tuning is described and allows easy adaptation to various facial expression analysis applications.

Keywords Emotion dynamics · Emotion recognition · Emotional surface · Kalman filtering · Simulation–optimization

1 Introduction

Person-to-person communication constitutes natural, highly dynamical and multimodal uncertain systems. Studies reveal that nonverbal components such as facial expressions, body language, prosody and intonation convey at least 65 % of the context information in a typical conversation [1]. Applications that strive to understand these communication modes and integrate them in the human-machine interfaces are crucial to "user centric experience" paradigms [2,3]. Although voice, face and gesture recognition are now used in video games and affective computing frameworks, the inference of emotional states remains as an open problem.

It has been demonstrated that recognizing emotions is not easy, even for humans, who employ specialized brain subsystems for the task [4]. Multimodal studies have shown that humans correctly recognize the conveyed emotion expressed through speech in about 60 % of interactions. For facial recognition, the success rate rises to 70–98 % [2,5,6]. This paper focuses on emotion recognition based on facial expressions. State-of-the-art reviews of automatic facial expression detection techniques can be found in [7] and [8].

As an introductory case, consider, as an example, the frames from a video, shown in Fig. 1 and the outputs from the commercially available edition of eMotion [9], in Fig. 2.

From eMotion's output data in Fig. 2, it would be impossible for a human subject to make an educated guess regarding the expressed emotion. If one performs the classification based solely on higher mean value, the result would be Sadness. However, watching the video, even without sound, a human would easily choose Anger as the emotional state of the speaker.

This work discusses a general model for the detection of emotional states and presents a model to detect slow dynamic emotions that constitute the perceived emotional state of the speaker. It is organized as follows: reference material is presented in Sect. 2, while Sect. 3 presents the general model, Sect. 4 describes the specific proposed model, the Kalman filtering technique and the heuristics used for model tuning, Sect. 5 describes the proposed experiment and results.

This is a revised and extended version of a paper that appeared at ENIA 2011, the Brazilian Meeting on Artificial Intelligence (http://www.dimap.ufrn.br/csbc2011/eventos/enia.php).

R. A. M. Gonçalves · D. R. Cueva · M. R. Pereira-Barretto (✉) · F. G. Cozman
Escola Politécnica, Universidade de São Paulo (USP), São Paulo, Brazil
e-mail: mrpbarre@usp.br

F. G. Cozman
e-mail: fgcozman@usp.br

Fig. 1 From *left* to *right*, eMotion classified these frames as happiness (100 %), sadness (70 %), fear (83 %) and anger (76 %), respectively. Video s43_an_2 of the eNTERFACE'05 Audio-Visual Emotion Database [26]. Extracted from [28]

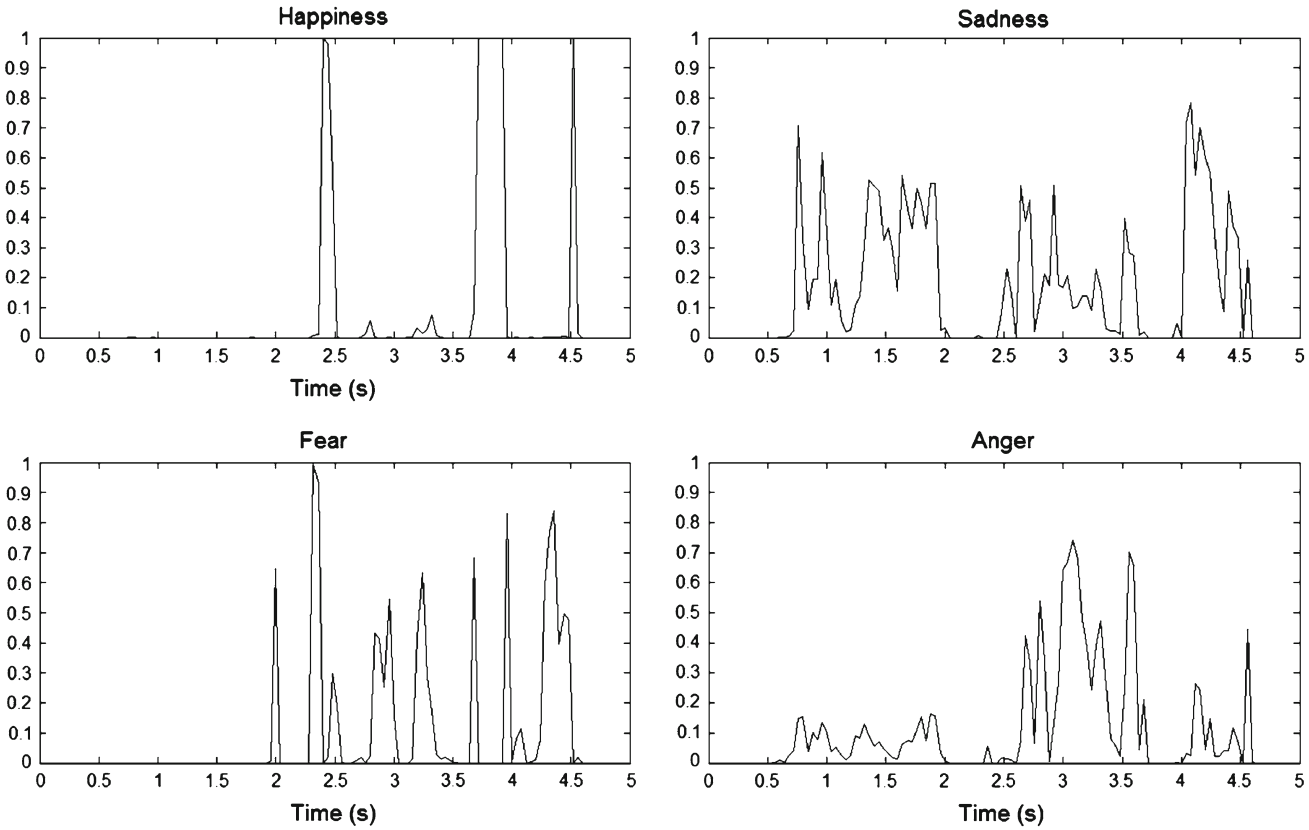

Fig. 2 Graphical representation of eMotion's output for the video of Fig. 1. eMotion analyses each video frame individually and outputs the estimated probability for each emotion category at that frame

2 Background

After decades of Behaviourism dominance in Psychology, Appraisal Theories gained strength since the 60's, [10,11]. These theories postulate that emotions are elicited from appraisals. Emotions, according to appraisal theorists, may be defined as "... an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" [10]. Appraisals differ from person to person but the appraisal processes are the same for all persons. Therefore, they offer a model which justifies a common behavior but, at the same time, allows for individual differences. From all events, the conveyed emotion, as perceived in facial expressions, is the focus of this work.

In the 70's, Ekman and co-workers proposed the universality of facial expressions related to emotions [6]. Their thesis was based on a series of experiments with different cultures around the world. Most notable were the results obtained with pre-literate and culturally isolated tribes which were able to classify photos of facial expressions better than chance [6]. A sample of their work is shown in Table 1, giving support for the universality of recognition of emotions on faces.

The 30-year long debate around the universality, its acceptance and its implications are discussed in [5] and [12].

Table 1 Median percentage agreement for forced choice

Culture group	Facial expression					
	Happy	Surprise	Sadness	Fear	Disgust	Anger
Western	96.4	87.5	80.5	77.5	82.6	81.2
Non-Western literate	89.2	79.2	76.0	65.0	65.0	63.0
Illiterate, isolated	92.0	36.0	52.0	46.0	29.0	56.0

Extracted from [5]

Ekman and Friesen also established the Facial Action Coding System (FACS), a seminal work for emotion recognition from faces by decomposing the face into AUs (Action Units) and assembling them together to characterize an emotional expression [13]. The universality thesis is strongly relevant to this work because it implies universality for the proposed model; the thesis, however, still receives criticism [14].

One could classify the recent approaches to computational facial expression analysis into two groups. In one group there are innovative techniques focusing on spatiotemporal features and usually employing classifiers based on HMM [15] and [16]. Their recent popularity due the arrival of cheap 3D cameras may lead to significant changes in this field. The second group consists of more traditional approaches: Haar-like and geometric features, polygonal and Bezier mesh fitting, Action Unit's tracking and energy displacement maps, [17–19] The later methods are currently employed in both academic and commercial developments and the most recent proposals employ multimodal analysis of emotional states [20].

Among the second group's most mature solutions, we cite eMotion, developed at Universiteit van Amsterdam [9], and FaceDetect, by the Fraunhofer Institute [21], both of which are commercially available. Both software packages focus on detecting emotion in facial expressions from each video frame, and they show excellent results in posed, semi-static situations. However, during a conversation, the face is distorted to speak in many ways, leading the algorithms to incorrectly detect the conveyed emotion. Even more, lip movement during a conversation, similar to a smile for instance, does not mean the speaker is happy. Instead, it may be an instantaneous emotion: the speaker saw something not related to the conversation, and that made him smile. There is a difference between the emotion expressed in the face and the general emotional state of the speaker.

3 Overview of proposed model

The proposed model to determine perceived emotion from instantaneous facial expressions is based on the displacement of a particle over a surface, subject to velocity changes proportional to the current probability of each emotion, at every moment. We propose calling this surface the "Dynamic Emotional Surface" (DES). Over the surface, attractors corresponding to each detectable emotion are placed. The particle moves freely over the DES; its velocity is at each instant proportional to the instantaneous emotions detected. The particle may also slide towards the neutral state, placed at the origin of the coordinate system, the point of minimum energy, or any other local minimum.

As input, the model takes emotion detection from video frames as worked by many authors [7,8,22,23]. Any of these software packages for facial expression analysis can be taken as a "raw sensor" from which data to be processed in the proposed model is obtained. Data are processed by Kalman filtering to remove noisy outputs and by an integration phase over a Dynamic Emotional Surface (DES), as depicted in Fig. 3.

Raw signals related to each emotion are fed into low-pass filters so both instantaneous marker expressions and erroneous high frequency variations are eliminated.

To illustrate this, consider a conversation with a friend: the overall conveyed emotion could be Happiness (the slow dynamic). But suddenly the speaker remembers someone he hates: Anger may be displayed as a marker expression. The event could be external: the speaker may see someone doing something wrong and may display Anger. In both cases, Anger is displayed as the fast dynamics, lasting no more than a couple of frames. For the listener, the appraisal process

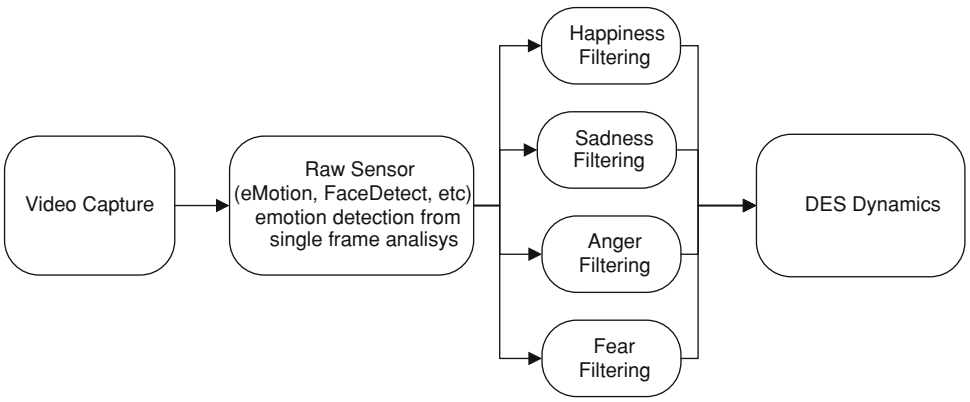

Fig. 3 Processing pipeline for the proposed model. The raw sensor output for each model's emotion is filtered individually with no prior knowledge of video's emotional content. The filtered outputs are applied to the integration stage over the DES

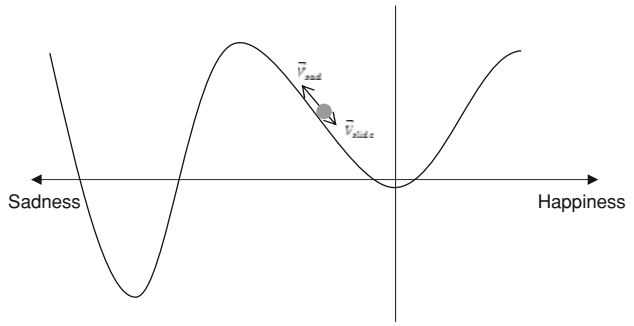

Fig. 4 An emotional curve. In this example the system detected an expression related to sadness, thus the particle has a $\overrightarrow{V}_{\text{sad}}$ component. The sliding velocity is represented as $\overrightarrow{V}_{\text{slide}}$ and it is proportional to the curve steepness, that is, tending to a stable point, normally neutral emotion

might lead to ignore Anger and continue the conversation, or to change the subject to investigate what caused this change in the speaker's face. The proposed model has been developed to detect the slow dynamic.

4 Proposed model

As stated before, the perceived emotion from instantaneous facial expressions is based on the displacement of a particle over a surface, subject to velocity changes proportional to the current probability of each emotion, at every moment, detected by raw sensors.

The instantaneous particle's velocity is determined by Eq. (1).

$$\overrightarrow{V}_{\text{p}} = \overrightarrow{V}_{\text{s}} + \sum_{a=1}^{N} \overrightarrow{V}_{\text{a}}, \tag{1}$$

where $\overrightarrow{V}_{\text{p}}$ particle velocity, $\overrightarrow{V}_{\text{s}}$ sliding velocity, parallel to DES' gradient at the current position, $\overrightarrow{V}_{\text{a}}$ velocity towards each attractor, always tangent to the DES.

Consider, as an example, the two-dimensional case where the detectable emotions are Happiness and Sadness, shown in Fig. 4.

The example demonstrates some key aspects of DES. The attractors for Happiness and Sadness are placed at $(\infty, 0)$ and $(-\infty, 0)$, respectively. When the raw sensor detects some probability or intensity of an emotion, this signal is interpreted as a velocity along the trajectory towards the correspondent attractor and the particle moves along the emotional curve. In the absence of emotional facial expressions, the particle slides to the local minimum. In this example, one may infer the emotional state of the speaker observing the position of the particle along the X axis.

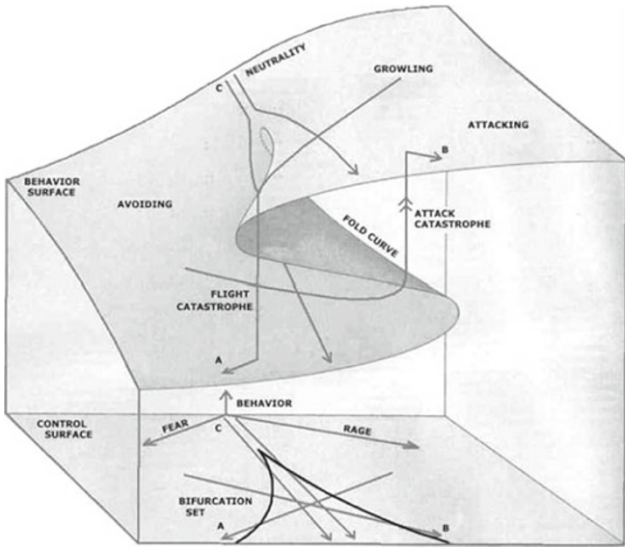

Fig. 5 Zeeman's emotional surface for the fight or flight case [24]

The DES concept extends this example by defining a surface or even a hypersurface over which attractors representing the modeled emotions are placed. The relationship between a particle's position and emotional classification is also defined. The idea of an emotional surface, as shown in Fig. 5 [11,24], has been proposed by psychologists to discuss someone's internal (appraised) emotion state trajectories; in this paper, it is used to detect the overall perceived emotion during a man-machine interaction.

The DES concept also differs from Zeeman's model on presenting the emotions as attractors positioned on the XY plane instead of attributing them to the axes themselves.

A DES in a 3D space is defined as Eq. (2).

$$\gamma(x, y) = (x, y, f(x, y)) \tag{2}$$

The velocity in the direction of each attractor, $\overrightarrow{V}_{\text{a}}$, is proportional to the probability of each emotion as detected by existing software such as eMotion and it is tangent to the surface. It is defined as Eq. (3).

$$V_{\text{a}} = F_{\text{a}} \frac{\nabla\gamma(x, y)}{|\nabla\gamma(x, y)|} \tag{3}$$

where F_{a} is the filtered signal associated with the attractor's emotion.

It should be noted that the frame-by-frame approach used by the raw sensors does not take into account the continuous natural facial movements and the transitions between expressions. As shown in Fig. 3, a filtering process is applied to raw sensor outputs prior to DES calculations.

The analysis of multimodal realistic videos must account for different noise sources in the process and its observation. Unexpected camera and head motions, face deformation due to speech, CCD performance and minor light source

variations result in intrinsically noisy data. Besides, low-pass filtering is necessary because the slow conveyed emotions are to be detected. Both Kalman filtering and moving-average filtering were tested, as presented in Sect. 5.3.

Due to these requirements, a Kalman filter is a natural candidate. Kalman filtering is a well-established technique for linear systems subject to zero mean Gaussian noise both in the process and the sensorial acquisition. There is no empirical evidence to support these hypotheses for the problem of emotional expression analysis. However, it was assumed, due to the complexity and apparent randomness of movements, that muscular facial deformations due to speech and light variations are in the scene. The rationale presented is, thus, the central limit theorem. Filtering convergence during the experiments gave further support for this assumption.

The use of Kalman filters requires the selection of underlying linear models for the update phase. It is proposed that a well-tuned first order system, as in Eqs. (4) and (5), doubles as the filter's internal update mechanism and low-pass filter. Filtering output for each emotion is described as F_a and used in Eq. (3).

$$\dot{x}_s = x_s, \tag{4}$$

$$F_\mathrm{a} = y = \frac{K x_s}{\tau}, \tag{5}$$

where K System's gain, τ System's time constant, x_s State variable, y Filter output.

The Kalman filtering equations are thus written as follows:

Predict:

$$x_{s,t} = x_{s,\,t-1}, \tag{6}$$

$$p = p + \frac{w}{\tau^2}, \tag{7}$$

where $x_{s,t}$ Current x value, $x_{s,\,t-1}$ x value in last instant estimation, w Covariance of the process noise, $N(0, w)$, p Covariance of x_t, $N(0,\ p)$.

Update:

$$m = \frac{\frac{pK}{\tau}}{p\left(\frac{K}{\tau}\right)^2 + v}, \tag{8}$$

$$x_{s,t} = x_{s,t} + m(r_t - y_t), \tag{9}$$

$$p = \left(1 - \frac{mK}{\tau}\right) p, \tag{10}$$

where m Residual covariance, v Covariance of observation noise, $N(0, v)$, r_t Current reading from facial expression analysis software, y_t Current filter output.

The estimation process has two steps. First, the filter runs prediction using a proper time step. If there is raw sensor information for that timestamp, it runs the update phase. One may notice that the state variable x_s represents only an internal calculated value. The proposed filtering relies only on readings from facial expression analysis software to calculate the internal state of the system.

Lastly we propose a simulation–optimization heuristic to tune system filters' w and v parameters. It employs Simulated Annealing (SA) to determine a set of parameters to minimize an energy function related to the error on classification. The simulation phase is comprised of a round of video analysis based on the current proposed parameters and is used to calculate a global energy value, the optimization phase is further discussed.

Defining vectors for process noise (Q_n) and observation noise (R_n) as follows:

$$Q_n = \left[w_\mathrm{happiness}, w_\mathrm{sadness}, w_\mathrm{anger}, w_\mathrm{fear}\right], \tag{11}$$

$$R_n = \left[v_\mathrm{happiness}, v_\mathrm{sadness}, v_\mathrm{anger}, v_\mathrm{fear}\right]. \tag{12}$$

Then defining a starting temperature (T_0) and a cooling constant $K_t < 1$:

$$T_{n+1} = K_t T_n. \tag{13}$$

The process iterates until the system's temperature matches room temperature (T_room). One may calculate the number of steps using Eq. (14):

$$N_\mathrm{SAsteps} = \mathrm{ceil}\left(\log_{K_t} \frac{T_0}{T_\mathrm{room}}\right). \tag{14}$$

For each video, the emotional particle's trajectory is divided in two halves. The energy (E_i) is calculated as the number of later half's points that are outside the sector of its nominal classification. A global energy measure is defined by Eq. (15).

$$E_{g,n} = \sum_{0}^{N_\mathrm{videos}} E_{i,n}. \tag{15}$$

The system then randomly generates neighbor parameter vectors Q_{n+1} and R_{n+1}. It reanalyzes the tuning videos and obtains $E_{\mathrm{global},n+1}$. The probability of accepting the new parameters as a solution is given by the Metropolis criteria:

$$P_\mathrm{Acceptance} = \min\left\{ \mathrm{e}^{\frac{1}{\frac{E_{g,n} - E_{g,n+1}}{T_{n+1}}}} \right. \tag{16}$$

These steps are summarized in Algorithm 1.

5 Experiments

Experiments were conducted to test the proposed model for the detection of the slow emotional dynamic.

5.1 Corpus selection

Selecting videos for emotion inference experiments presents some challenges: the videos must respect the conditions

```
Choose T_0,T_room, K
Calculate N_steps
Initialize Q, R randomically
Q_new ← Q
R_new ← R
E ← E_best ← MAX_INT
For I = 1, N_steps:
   E_global ← 0
   For J = 1, N_videos:
     Calculate E_v by simulation, using Q_new,R_new
     E_global ← E_global + E_v
   End
   Calculate P
   If accepted:
     E ← E_global
     Q ← Q_new
     R ← R_new
   End
   If E_global < E_best:
     E_best ← E_global
     Q_best ← Q_new
     R_best ← R_new
   End
   Select randomically new value for randomically selected component of Q,R _
   and generate Q_new, R_new
End
```

Algorithm 1 Simulation-optimization algorithm for tuning filter's parameters

imposed by the raw sensors such as lighting, head positioning, duration and resolution, and they must also contain images with expressions in a natural way. Additionally, they must be generally available, so further research may reproduce and compare results.

The eNTERFACE'05 Audio-Visual Emotion Database [25] was selected as baseline corpus for both the research on emotion inference from facial expressions and multimodal inference [26]. This database consists of volunteers acting in a series of short scenes, expressing emotions through facial expressions, speech and vocalization. The volunteers are not professional actors and, as it will be demonstrated, there are some cases where is not possible to classify the conveyed emotion based solely on the facial expressions. Therefore, an initial experiment was conducted to select viable videos.

A set of 50 videos from the eNTERFACE'05 Audio-Visual Emotion Database has been selected. These videos were presented twice, one at a time, without sound, to 17 undergraduate subjects from the Mechatronics course. The students were given a multiple choice formulary where they were asked to classify each video as Happiness, Sadness, Anger or Fear, leaving no blanks. This methodology differs from [27] and [28] where the videos were chosen by the researchers only. Experimental results are shown in Tables 2, 3, 4 and 5 and in Fig. 6.

Table 2 Human classification for videos classified as happiness

File	Happiness	Sadness (%)	Anger (%)	Fear (%)
s2_ha_2	100.0	0.0	0.0	0.0
s4_ha_2	100.0	0.0	0.0	0.0
s4_ha_4	100.0	0.0	0.0	0.0
s12_ha_3	100.0	0.0	0.0	0.0
s25_ha_2	94.1	0.0	5.9	0.0
s29_ha_3	94.1	5.9	0.0	0.0

These videos were then categorized as valid emotional samples or not, based on an agreement score of at least 90 % of the expected values shown in Table 1. The minimum scores were thus 86.8, 69.8, 73.1 and 72.5 %, yielding 31 valid videos: 7 for Happiness, 6 for Fear, 8 for Anger and 10 Sadness.

5.2 Data acquisition

This section describes the data acquisition specifically related to the eMotion software. The process starts by splitting

Table 3 Human classification for videos classified as fear

File	Happiness (%)	Sadness (%)	Anger (%)	Fear (%)
s2_fe_4	6.3	37.5	25.0	31.3
s14_fe_2	0.0	35.3	52.9	11.8
s24_fe_3	11.8	5.9	0.0	82.4
s24_fe_4	0.0	23.5	0.0	76.5
s25_fe_2	5.9	0.0	11.8	82.4
s28_fe_2	5.9	5.9	82.4	5.9
s33_fe_5	0.0	5.9	47.1	47.1
s36_fe_2	0.0	23.5	5.9	70.6
s37_fe_3	5.9	11.8	47.1	35.3
s38_fe_3	0.0	17.6	5.9	76.5
s42_fe_1	0.0	0.0	47.1	52.9
s43_fe_2	0.0	0.0	0.0	100.0

Table 4 Human classification for videos classified as anger

File	Happiness (%)	Sadness (%)	Anger (%)	Fear (%)
s2_an_2	31.3	6.3	43.8	18.8
s4_an_2	0.0	23.5	76.5	0.0
s4_an_5	0.0	0.0	76.5	23.5
s14_an_1	0.0	0.0	88.2	11.8
s25_an_2	5.9	17.6	52.9	23.5
s28_an_4	0.0	70.6	29.4	0.0
s29_an_2	94.1	0.0	5.9	0.0
s29_an_4	70.6	17.6	5.9	5.9
s33_an_2	6.3	25.0	56.3	12.5
s36_an_3	11.8	35.3	35.3	17.6
s37_an_1	11.8	47.1	35.3	5.9
s38_an_1	0.0	0.0	88.2	11.8
s43_an_2	0.0	0.0	100.0	0.0
s43_an_3	0.0	0.0	94.1	5.9
s43_an_4	0.0	0.0	100.0	0.0
s43_an_5	0.0	0.0	100.0	0.0
s44_an_4	0.0	0.0	70.6	29.4

the selected videos, according to the criteria in Sect. 5.1, into two groups: one for system tuning and one for testing. Each video has been submitted sequentially to the eMotion software and control points for mesh adjustment were selected. After mesh fitting, each video has been played back, observing if the mesh remains attached to face's control points during the whole video. In case of abnormal mesh deformation, the current analysis was discarded and the operator had to return to the mesh fitting step.

The output data for each video has been collected in a separated CSV dump file containing frame-by-frame values.

Table 5 Human classification for videos classified as sadness

File	Happiness (%)	Sadness (%)	Anger (%)	Fear (%)
s1_sa_1	0.0	25.0	18.8	56.3
s2_sa_4	0.0	82.4	11.8	5.9
s4_sa_1	0.0	94.1	0.0	5.9
s14_sa_3	5.9	70.6	11.8	11.8
s14_sa_5	70.6	5.9	11.8	11.8
s29_sa_1	0.0	82.4	11.8	5.9
s29_sa_3	0.0	64.7	5.9	29.4
s33_sa_2	0.0	82.4	5.9	11.8
s36_sa_2	0.0	88.2	0.0	11.8
s42_sa_1	0.0	88.2	11.8	0.0
s43_sa_1	0.0	100.0	0.0	0.0
s43_sa_3	0.0	94.1	0.0	5.9
s43_sa_4	0.0	100.0	0.0	0.0
s43_sa_5	0.0	94.1	0.0	5.9

5.3 Filter selection

The results of Kalman filtering and moving-average (window size of 20 frames) for the example video (sample frames in Fig. 1 and raw sensor output on Fig. 2) are shown in Fig. 7.

As it can be seen from Table 6, the overall emotion conveyed by the video, Anger, has been correctly detected with Kalman filtering, although with a large standard deviation. Kalman filtering was therefore selected to conduct automatic classification.

5.4 DES selection

A paraboloid with parameters shown in Eq. (17) and attractors placed as in Table 7 has been chosen for DES.

$$\gamma(x, y) = (x, y, a_1 x^2 + a_2 y^2) \tag{17}$$

$$a_1 = a_2 = 0,\ 6.$$

One may note that the Fear attractor was placed in the fourth quadrant, which is not the usual position on the Arousal-Valence field. In fact, the placement of the attractors is arbitrary and depends on the DES, the phenomena to be modeled and how one defines the classifying function. The paraboloid DES was used to model "reasonable" social displays of emotion and the particle's position is said to be related to one of the attractors if in the same quadrant. It also yields to simplifications as follows.

Considering $\overrightarrow{\mathrm{P}}$ as the particle's current position and $\overrightarrow{\mathrm{A}}$ the position of the attractor (emotion), their distance can be calculated as Eq. (18).

$$\overrightarrow{AP} = \overrightarrow{A} - \overrightarrow{P} = \left[a_{px},\ a_{py},\ a_{p\sigma}\right]. \tag{18}$$

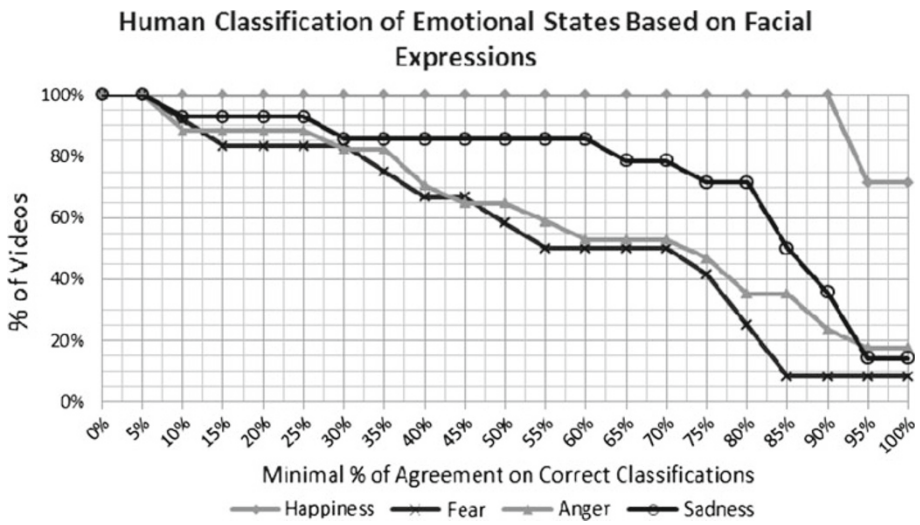

Fig. 6 Human classification of emotional states based on facial expressions

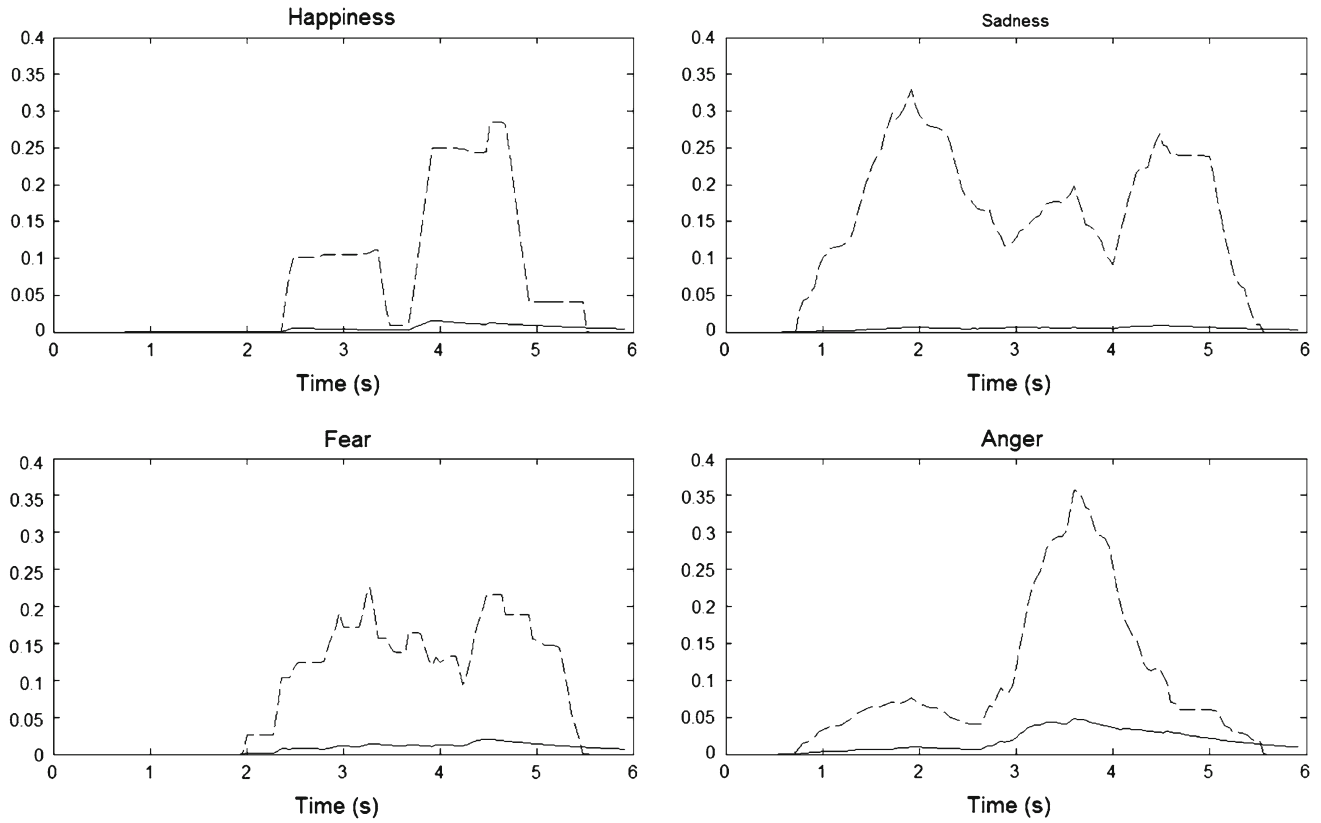

Fig. 7 Moving average (*dashed*) and proposed Kalman filtering (*solid*) outputs for example video on Fig. 1

If we define a ratio r as in Eq. (19), DES $S(x)$ may be written as a function of the variable x as

$$r = \left| \frac{a_{py}}{a_{px}} \right|, \; a_{px} \neq 0, \tag{19}$$

$$S(x) = \gamma(x, rx). \tag{20}$$

The particle's velocity is calculated as

$$V_{\mathrm{a}} = F_{\mathrm{a}} \frac{\left[1, r, 2(a_1 + a_2 r^2) * P_x\right]}{\sqrt{1 + r^2 + \left[2(a_1 + a_2 r^2) P_x\right]^2}} \tag{21}$$

Figure 8 shows the XY projection of the emotional particle's trajectory for the example video (all frames).

The XY projection of the emotional particle's trajectory for the example video reveals that the emotional state of the speaker may be described as Anger, as the particle moves

Table 6 Comparison between unfiltered signals, moving average and proposed Kalman filtering

Emotion	Original		Moving average		Kalman	
	μ	Σ	μ	Σ	μ	σ
Happiness	0.175	0.634	0.175	0.237	0.131	0.137
Sadness	0.377	0.532	0.377	0.254	0.139	0.073
Fear	0.211	0.544	0.211	0.206	0.219	0.187
Anger	0.236	0.434	0.236	0.257	0.511	0.423

Table 7 Attractor placement

Emotion	Attractor projection
Happiness	$[\infty, \infty, 0]$
Anger	$[-\infty, \infty, 0]$
Sadness	$[-\infty, -\infty, 0]$
Fear	$[\infty, -\infty, 0]$

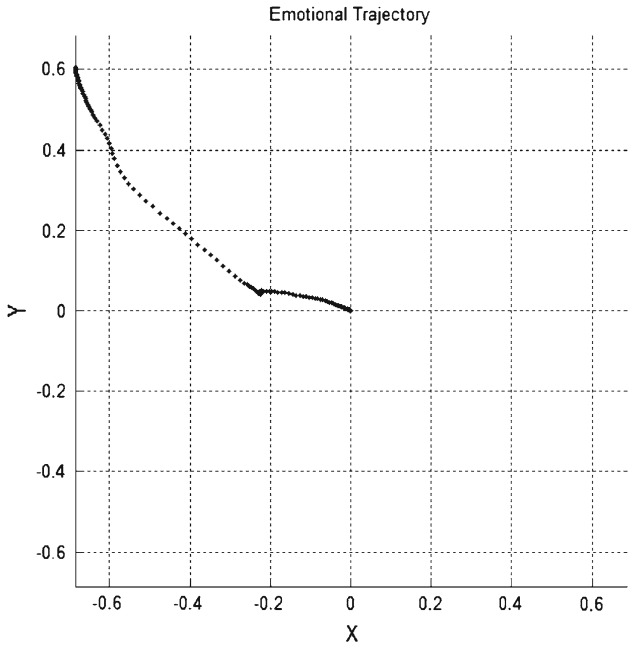

Fig. 8 Projection of the emotional trajectory for all frames on sample video (some frames on Fig. 1). Each *dot* represents the best estimate of the subject's emotional state at each frame

on the second quadrant. This inference corresponds to the human observation; see Table 10, "s43_an_2".

5.5 Tuning Kalman filters

The 31 valid videos were split in two groups: 16 videos for Kalman filter tuning and 15 for testing the proposed model.

Based on previous experience in system tuning [27,28], system gain and time constant for all underlying linear models were fixed for all four filters. Algorithm 1 was used to calibrate w and v parameters. The initial w and v were chosen randomly from a uniform distribution in the interval [0.001, 1000]. Additional starting conditions were:

$$T_0 = 2,500.00,$$
$$T_{\text{room}} = 10,$$
$$K_t = 0.9995.$$

These conditions lead to 11,041 iterations. Tuning was repeated for 18 runs, looking for convergence to a minimum. The results are presented in Table 8.

The graph in Fig. 9 represents all accepted solutions during the simulation-optimization process that resulted in 447 as minimum energy.

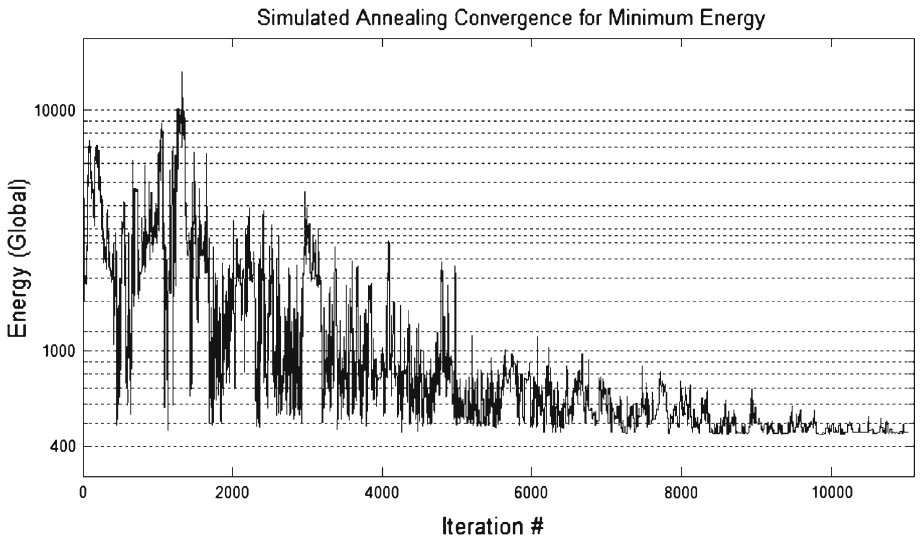

Fig. 9 Convergence for the best solution obtained using the proposed simulation-optimization method

Table 8 Simulated annealing results, 18 runs with 11,041 iterations each

$E_{g,\text{minimum}}$					
447	452	459	471	478	481
481	485	498	540	546	4,575
4,575	4,618	5,862	5,998	6,124	6,147

Table 9 Kalman filtering parameters for eMotion as raw sensor

	w	v	K	τ
Happiness	207.91	692.04	5	1.5
Anger	79.16	558.61	5	1.5
Sadness	270.90	631.64	5	1.5
Fear	490.95	483.38	5	1.5

Table 10 Comparison between human evaluation and the proposed Kalman filtering with DES algorithm

File	Classifications	
	Human	System
s2_ha_2	Happiness	Happiness
s25_fe_2	Fear	Fear
s29_ha_3	Happiness	Happiness
s38_an_1	Anger	Anger
s38_fe_3	Fear	Fear
s42_sa_1	Sadness	Sadness
s43_an_2	Anger	Anger
s43_an_3	Anger	Anger
s43_an_4	Anger	Anger
s43_fe_2	Fear	Fear
s43_ha_1	Happiness	Happiness
s43_sa_1	Sadness	Sadness
s43_sa_3	Sadness	Sadness
s43_sa_4	Sadness	Sadness
s43_sa_5	Sadness	Anger

The resulting parameters are presented in Table 9 along with the defined gains and time constants.

5.6 Automatic classification

The 15 remaining videos, i.e., those not used for adjusting the Kalman filter, were then submitted to the system, yielding the results shown in Table 10.

The XY projection for (misclassified) file s43_sa_5 is shown in Fig. 10.

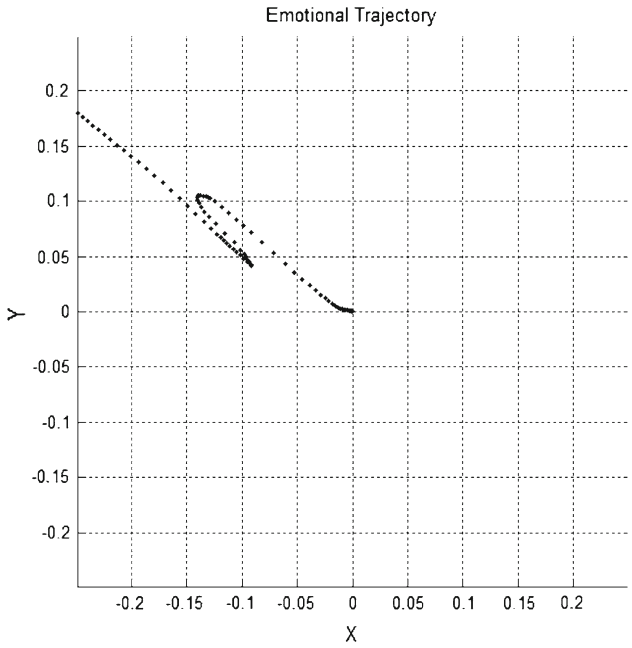

Fig. 10 Emotional trajectory for file "s43_sa_5". Note that the particle oscillates inside the second quadrant yielding the classification as Anger. The correct classification is Sadness

6 Conclusions

A reference model for recognition of emotions on faces has been introduced, as well as a computational model to detect slow conveyed emotions and to infer the speaker's overall emotional state. The model was tested and presented excellent results.

The proposed architecture allows these techniques to be integrated with almost any facial analysis expression software available with minimal changes. The proposed simulation-optimization heuristic leads to automatic configuration and system tuning. One should note that although there are recent techniques that employ spatiotemporal features, they could still benefit from the proposed model to infer general perceived emotions in natural interactions.

In future work we plan to test the model for fast emotions. The main obstacle we foresee is the lack of a corpus for this kind of test. Finally, we plan to apply the proposed model in a multimodal inference engine, as proposed in [28].

Aclnowledgments The authors thank CNPq and FAPESP (project 2008/03995-5), for their financial support.

References

1. Birdwhistell R (1970) Kinesics and context. University of Pennsylvania Press, Philadelphia
2. Picard RW (1995) Affective computing. MIT Press, Cambridge

3. Picard R (2003) Affective computing: challenges. Int J Hum Comput Stud 59:55–64
4. Brothers L (1999) Emotion and the human brain. In: The MIT encyclopedia of the cognitive sciences. MIT Press, Cambridge, pp 271–273
5. Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull 115(1):102–141
6. Ekman P, Friesen WV, Ellsworth P (1972) Emotion in the human face. Pergamon Press, Oxford
7. Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 22(12):1424–1445
8. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31:39–58
9. Azcarate A, Hageloh F, Sande K, Valenti R (2005) Automatic facial emotion recognition. Universiteit van Amsterdam
10. Scherer KR (2001) Appraisal considered as a process of multilevel sequential checking. In: Appraisal processes in emotion: theory, methods, research. Oxford University Press, Oxford, pp 92–120
11. Sander D, Grandjean D, Scherer KR (2005) A systems approach to appraisal mechanisms in emotion. Neural Netw 18:317–352
12. Ekman P (1993) Facial expression and emotion. Am Psychol 48(4):376–379
13. Ekman P, Friesen WV (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Washington
14. Naab PJ, Russell JA (2007) Judgments of emotion from spontaneous facial expressions of New Guineans. Emotion 7(4):736–744
15. Le V, Tang H (2011) Expression recognition from 3D dynamic faces using robust spatio-temporal shape features. In: Proceedings of the IEEE international conference on automatic face and gesture recognition and workshops (FG 2011), pp 414–421
16. Sun Y, Yin L (2008) Facial expression recognition based on 3D dynamic range model sequences. In: Proceedings of the 10th European conference on computer vision: part II, pp 58–71
17. Valstar M, Gunes H (2007) How to distinguish posed from spontaneous smiles using geometric features. In: Proceedings of the ICMI'07: 9th international conference on multimodal interfaces
18. Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Measuring facial expressions by computer image analysis. Psychophysiology 36(2):253–263
19. Essa IA, Pentland AP (1997) Coding, analysis, interpretation, and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 19(7):757–763
20. Milanova M, Sirakov N (2008) Recognition of emotional states in natural human–computer interaction. In: Proceedings of the ISSPIT 2008-IEEE international symposium on signal processing and information technology, pp 186–191
21. Fraunhofer(IIS): Fraunhofer Facedetect. http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo/
22. Tian Y-l, Kanade T, Cohn JF (1978) Facial expression analysis. In: Handbook of facial recognition. Springer, Berlin
23. Fasel B, Luettin J (2003) Automatic facial expression analysis: a survey. Pattern Recogn 36:259–275
24. Zeeman EC (1976) Catastrophe theory. Sci Am 234(4):65–83
25. Martin O, Kotsia I, Macq B, Pitas I (2006) The eNTERFACE 05 audio-visual emotion database. In: Proceedings of the 22nd international conference on data engineering workshops, pp 8–8
26. Cueva DR, Gonçalves RAM, Pereira-Barretto MR, Cozman FG (2011) Fusão de Observações Afetivas em Cenários Realistas (in Portuguese). Anais do XXXI CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO—Encontro Nacional de Inteligência Artificial, pp 833–842
27. Gonçalves RAM, Cueva DR, Pereira-Barretto MR, Cozman FG (2011) Determinação da Emoção Demonstrada pelo Interlocutor (in Portuguese). Anais do XXXI CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO—Encontro Nacional de Inteligência Artificial, pp 737–748
28. Cueva DR, Gonçalves RAM, Cozman FG, Pereira-Barretto MR (2011) Crawling to improve multimodal emotion detection. Lecture Notes in Artificial Intelligence, vol 7094, Part II. Springer, Berlin, pp 343–350