A Model for Facial Emotion Inference Based on Planar Dynamic Emotional Surfaces

Ruivo, J. P. P.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Negreiros, T.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Barretto, M. R. P.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Tinen, B.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

***Abstract*—Emotions have direct influence on the human life and are of great importance in relationships and in the way interactions between individuals develop. Because of this, they are also important for the development of human-machine interfaces that aim to maintain a natural and friendly interaction with its users. In the development of social robots, which this work aims for, a suitable interpretation of the emotional state of the person interacting with the social robot is indispensable. The focus of this paper is the development of a mathematical model for recognizing emotional facial expressions in a sequence of frames. Firstly, a face tracker algorithm is used to find and keep track of faces in images; then the found faces are fed into the model developed in this work, which consists of an instantaneous emotional expression classifier, a Kalman filter and a dynamic classifier that gives the final output of the model.**

Keywords—emotion recognition, facial emotion, Kalman filter, machine learning

I. Introduction

Emotions influence the human behavior and the way individuals interact and relate to other members of society. They permeate one's daily life and determine how people react to the various situations they encounter in their routines.

Studies indicate that people with impairments to express or recognize feelings end up having great difficulty keeping even casual relationships [1]. Emotions also help the body prepare for specific external events. For example, the fear people may experience when they see a large object coming fastly towards them stimulates blood circulation in their legs, allowing them to act promptly and respond trying to avoid the object.

Computer interfaces that can understand the emotional state of its users can communicate more naturally compared to interfaces without this capability. Affective computing comes to deal with the integration of the concept of emotion in the computational area [2].

Emotions are characterized by signs in voice, speech and body movements, which are recognized regardless of culture, possibly being a legacy of human evolution and not a result of personal experiences of the individual [3]. Particularly in the face, the most obvious signs are presented in the regions of the mouth, eyes and eyebrows. Ekman and Friesen showed evidence for the hypothesis of universality of emotional facial expressions in intercultural studies with illiterate populations of Papua New Guinea and investigated the influence of the cultural phenomena [3].

Works from Ekman [4] [5] propose the existence of six major universal emotions: joy, sadness, surprise, fear, anger and disgust. An emotional display can either be classified as belonging to one category, such as joy, or more than one category, forming composite emotions, such as the mixture of fear and angry, or joy and surprise.

This study aims to identify five basic emotional states: Happiness, Sadness, Anger and Fear, plus the Neutral state, which could be understood as the absence of emotions. The model proposed in this work does not try to describe short-lived or rapidly changing emotions (micro expressions, in the works of Ekman), but focuses on trying to detect lasting emotional states people may be subject to. The dynamic model for emotion recognition presented in this work is a novel model based on the work of [6].

The rest of this paper is organized as follows. In Section II are reviewed previous works regarding automatic emotion inference. The adopted methodology, including face detection, feature extraction, instantaneous emotion recognition and dynamic emotion recognition is then presented at Section III. Section IV describes the results obtained. Finally, the conclusions and future work directions are presented on Section V.

II. Bibliographic Review

There are three main approaches to emotions classification: discrete model, dimensional model and the approach based on evaluation mechanisms [7].

The discrete model arranges emotions in categories, like the basic emotions of Ekman. Categorization of emotions is an intuitive and practical way to identify them, even if a large number of classes is necessary in order to classify all of the known affective states. Many of the works developed in the area utilizes this approach [8] [7] [9] [10].

The dimensional model seeks to describe the emotions by means of some criteria or dimensions. Two key dimensions are valence and arousal [10] [11]. Valence transmits how the person feels under the influence of a certain emotion, and can assume continuous values ranging from extreme sadness, for negative valence, to extreme happiness, for positive valence. Arousal is associated to the possibility of an individual to take or to perform an action under influence of an emotion, and can assume continuous values ranging from an extremely passive attitude, for negative arousal, to an extremely active attitude, for positive arousal. Some authors [12] suggest other dimensions for the model, such as dominance. Dominance is related to the control somenone has over a situation while under the influence of an emotion, and can assume continuous values ranging from total lack of control to total control of the situation. The dimensional model avoids the need for an extensive list of categories. Emotions are identified depending on its position on the model's axes. However, because of the limited number of dimensions this approach deals with, the projection of an emotion to the model's axes could cause loss of information [7].

The evaluation approach classifies emotional displays based on a set of assessments of the event that caused such display. For a given emotion, it is evaluated how relevant it is the event that elicited the emotion, what are its implications, the individual's ability to deal with these implications, and what is the significance of that event for the society the individual inhabits [13]. This approach is less simple and intuitive when compared to the others, as it requires a detailed analysis of the situations that elicited the emotions.

Pantic [14] suggests automatic recognition of facial emotion expressions to be done in three main steps: face detection, extraction of relevant features of the face and emotion classification

Face detection is a crucial step in the recognition of expressed emotions, and comprises of locating faces in still images or image sequences. In several works, such images are obtained under conditions that helps face detection algorithms, like the capture of the face in frontal orientation, without occlusions, and under uniform lighting conditions. However, in real situations, these conditions rarely can be reproduced, which makes the problem more challenging. Consequently, an ideal method of facial detection should deal with problems such as the different scales and orientations the human face may take, besides having to consider possible partial occlusions of the face and changes in the lighting conditions.

Extraction of face relevant characteristics has the purpose of generating a feature vector to be used for the emotion identification. It seeks to describe the face through certain categorical or numerical information that should contribute to the recognition of the emotional state of the analyzed person. These characteristics may be based on features of the human face such as eyebrows, nose and mouth, or may be based on mathematical models. These models, in turn, may follow an analytical approach, in which the face is represented by a set of points or patterns of interest that contain specific regions of the face; or they may follow a holistic approach, in which the face is seen as a unit, with its particular shape and texture. Hybrid approaches also exist, in which features of the two above-mentioned approaches are combined. Different scales and orientations of the face, as well as partial occlusion and noise, hamper the execution of this step.

The extracted features vector should then be used to estimate the expressed emotion via a classification algorithm. In this step, any of the approaches presented for emotion classification may be used; however, much of the work done in the area uses the discrete approach [15]. The classification of the facial emotion expressions is done by machine learning algorithms trained with the feature vectors extracted from the samples of one or more training databases. Examples of these algorithms are Support Vector Machines (SVMs), Decision Trees and Neural Networks (NNs).

The present work introduces a fourth step to the process proposed by [14] and includes the usage of a continuous emotional classifier model, following the line of work of [9] and [16]. This step was introduced so that the model would be able to detect long-lasting emotional states rather than instantaneous emotional displays; also, it should help with the minimization of the influences of natural noises, like laugh and speech, that deform the face and difficult the determination of someone's facial emotion expression.

Figure 1 presents a flow diagram of the steps aforementioned.

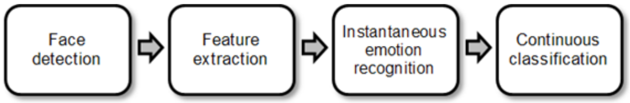

Fig. 1: Flow diagram for the proposed model.

One way to describe one's facial emotion expression is to use the Facial Actions Coding System (FACS) [17]. This system defines 44 Action Units (AU), each one representing the facial movements caused by muscle activity in a specific region of the face. Studies show that a particular subset of 15 of these AUs have greater relevance in the communication between humans [18].

FACS can be understood as an abstraction layer of the underlying facial muscle activity. Through the identification of the level of activity of the relevant AUs, one can infer the related muscles' activities and the corresponding facial expression. FACS defines, for example, involuntary and sincere expression of happiness as the activation of AUs numbers 6 and 12, that is, the lifting of the cheeks and the lateral and vertical extent of the lips, respectively. A forced (faked) expression shows only activation of AU 12 instead. This differentiation is possible because AU 12, which is the contraction of the zygomatic major muscle, is voluntary, while AU 6, contraction of the orbicularis oculi muscle, occurs involuntarily.

Furthermore, FACS brings into consideration the duration and intensity of AUs. Spontaneous muscle activations are in the range 250ms to 5 seconds, depending on the AU [19]. Rules for determining the intensity of each AU are also determined on FACS, for example, as the degree of elevation of the corner of the lips to the AU 12 or the wrinkle density over the nose for AU 44.

As noted in [9] [20], two different categories of properties could be extracted from faces: geometric properties and appearance properties.

Methods based on geometric properties look for characteristic regions of the face, such as eye contour, representing the shape and geometry of the features to be studied. For the extraction of data in video, one approach is the optical flow, as in [21], with tracking of characteristic points. Another approach are three-dimensional methods [22], which were developed along with the development of three-dimensional videos. In the solution presented in [23], the Active Shape

Model [24] and a Kalman filter were used to locate specific areas such as mouth and eyes in each frame of a video.

Appearance-based methods, however, search for changes in texture, such as wrinkles on the face. These methods can be used to describe the whole face or specific regions of interest [25] [18].

Following Figure 1, the next step, emotion classification, can be based on neural networks (NN), support vector machines (SVM) or hidden Markov chains (HMM) [9] [22] [22] [26], among other algorithms [18] [23].

It should also be noted that the humans' emotions detection system is not perfect, and emotions are not always interpreted correctly [14]. Donato [21] shows that people who had no training were able to correctly identify emotions in about 80% of a set of photos, but trained people, such as those passing throught FACS training, have a hit rate of about 90%. For Russell [27], however, a number of studies show that the rate of recognition by individuals varies according to the experimental conditions, ranging from about 55% to about 95%; also, negative emotions, such as anger and sadness, have a significantly lower accuracy recognition rate than positive ones.

The instantaneous emotion recognition model presented in this work is based on the work of Loconsole et al. [28]. In the referred work, an emotion classifier (namely, a random forest) based on geometric facial features is trained and used to differentiate images of faces expressing five emotional states: Joy, Sorrow, Surprise, Fear, Disgust and Anger. The authors analyze the accuracy their model achieved with and without calibration with neutral faces and considering different quantities of learned facial expressions. Also, they compare the accuracy of their model with that of other authors' models, and conclude their model achieved higher accuracy for the experiments made.

III. METHODOLOGY

This section briefly introduces the methods and techniques used to implement each of the steps shown in the diagram of Figure 1.

A. Face Detection

In this step, the Chehra Face Tracker is used [29]. This tracker detects and keeps track of faces in input images. It can be classified as a discriminative tracker, as it uses facial landmarks and discriminative functions to describe the current state of the face of a person, rather than a generative tracker, which would seek parameters that would maximize the probability of the deformable model to reconstruct a given face [29].

The Chehra Face Tracker uses an incremental parallel cascade of linear regressions to train the model, which has a better performance on face tracking in videos when compared to both the parallel cascade of linear regressions and the sequential cascade of linear regressions, showing better adaptation over time and robustness to environment changes on the face [29].

The tracker is capable of handling new training samples without having to retrain the model from scratch. It can also automatically tailor the model to the subject being tracked and to the imaging conditions, hence becoming person-specific over time [29].

B. Feature Extraction

Once the face tracker is able to fit the face model on one of the found faces in the image, one can proceed to extract features of interest from it.

The process of choosing what features to extract is not trivial, as the chosen feature set should be one that describes the studied concepts (in this case, the five facial emotion expressions: Happiness, Sadness, Anger and Fear, plus the Neutral state), so the trained classifier may have a better chance of learning how to properly differentiate amongst samples of these concepts. Loconsole [28] presents a feature set which is intended to differentiating among facial displays of Ekman's six basic emotions. This set comprises of two kinds of features: linear features and eccentricity features. While the linear features are determined by calculating the normalized linear distances between two given landmarks outputted by the face tracking model, the eccentricity features are given by the eccentricity measures of ellipses fitted over groups of three facial landmarks.

In the present work, Loconsole feature set is adopted with some new features added to it. The added features were chosen based on facial cues Ekman found to be of relevance in the process of facial emotion recognition [4]. The complete set of features adopted is described in Table I (refer to Figure 2 for the landmark's labels referenced in the table).

Table I: Extracted feature set

Name	**Measure**	**By**
F1	$\overline{UEBl_{m7y}UEl_{m3y}}/DEN$	[28]
F2	$\overline{U_{m1y}SN_y}/DEN$	[28]
F3	$\overline{D_{m2y}SN_y}/DEN$	[28]
F4	$\overline{EBlr_{Mx}EBrl_{Mx}}/DEN$	Us
F5	$\overline{A_{My}D_{m2y}}/DEN$	Us
F6	$\overline{B_{My}D_{m2y}}/DEN$	Us
F7	$\overline{A_{My}U_{m1y}}/DEN$	Us
F8	$\overline{B_{My}U_{m1y}}/DEN$	Us
F9	$\overline{EBlr_{My}Elr_{My}}/DEN$	Us
F10	$\overline{EBrl_{My}Er\text{ç}_{My}}/DEN$	Us
F11	$\angle(A_m, D_{m2}, B_m)$	Us
F12	$\angle(A_M, U_{m1}, B_M)$	Us
F14	$\angle(EBll_M, EBl_{aux}, EBlr_M)$	Us
F13	$\angle(EBrr_M, EBr_{aux}, EBrl_M)$	Us
F15	$\angle(EBllm_m, EBlr_M, EBl_{aux})$	Us
F16	$\angle(EBrr_M, EBrl_M, EBr_{aux})$	Us
F17	$Ecc(A_M, B_M, D_{m2})$	[28]
F18	$Ecc(A_M, B_M, D_{m2})$	[28]
F19	$Ecc(Ell_M, Elr_M, UEl_{m3})$	[28]
F20	$Ecc(Ell_M, Err_M, DEl_{m4})$	[28]
F21	$Ecc(Erl_M, Err_M, UEr_{mr})$	[28]
F22	$Ecc(Erl_M, Err_M, UEr_{m6})$	[28]
F23	$Ecc(EBll_M, EBlr_M, UEBl_{m7})$	[28]
F24	$Ecc(EBrl_M, EBrr_M, UEBr_{m8})$	[28]

In Table I, $\overline{(P_1P_2)}$ represents the linear distance between points P_1 and P_2, and the indices x and y are used to represent the horizontal and vertical points' coordinates, respectively. The notation $\angle(P_1, P_2, P_3)$ represents the internal angle between points P_1, P_2 and P_3, in radians. Finally, $Ecc(P_1, P_2, P_3)$ represents the eccentricity of an ellipse fitted over the points P_1, P_2 and P_3. The measure of eccentricity of an ellipse is given by the formula below (refer to Figure 3).

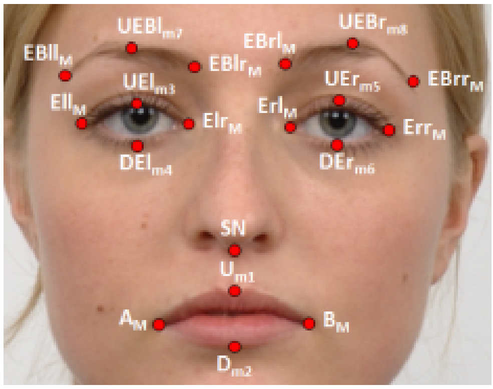

Fig. 2: The facial landmarks considered for the feature extraction process (taken from [28]).

$$Ecc(P_1, P_2, P_3) = \sqrt{\frac{\left(\frac{P_{1x} - P_{3x}}{2}\right)^2 + \left(\frac{P_{1y} - P_{2y}}{1}\right)^2}{\left(\frac{P_{1x} - P_{3x}}{2}\right)^2}} \tag{1}$$

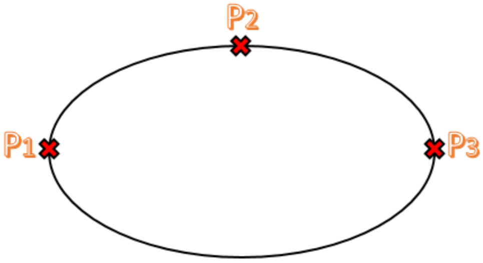

Fig. 3: An ellipse and the necessary points to the calculation of its eccentricity.

Feature F4 is a measure of the horizontal distance between the inner points of the eyebrows. This distance should be smaller in angry faces (which usually present the inner points of the eyebrows closer together) and bigger in surprised faces (which usually present the inner points of the eyebrows farther apart), for example.

Features F5 and F6 are measures of the vertical distances between the leftmost and the rightmost points of the mouth and the bottommost point of the mouth, respectively. These features should be helpful in differentiating facial expressions that present open mouths (like an angry expression, with exposed teeth) and closed mouths (like in a neutral expression). Also, they should be helpful in detecting if the analysed face is currently speaking or not.

Features F7 and F8 are similar to F5 and F6, but measure the vertical distances between the leftmost and the rightmost points of the mouth and the topmost point of the mouth. They have the same purpose features F5 and F6 have.

Features F9 and F10 are measures of the vertical distance between the inner points of the eyebrows and the inner points of the eyes. These features should help to differentiate facial expressions that present the inner corners of the eyebrows lifted (like in a surprised expression) from facial expressions that present the inner corners of the eyebrows lowered (like in an angered expression).

Feature F11 is the measure of the inner angle formed by the leftmost and rightmost points of the mouth with the bottommost point of the mouth. Feature F12 is the measure of the inner angle formed by the leftmost and the rightmost points of the mouth with the topmost point of the mouth. Together, they should be helpful in describing if the mouth is closed of opened, similarly to the features F5 to F8.

Features F13 and F14 are the measures of the inner angles formed by the corner of the eyebrows with the central point of each eyebrow. They should be helpful in describing if the eyebrows are arched (like in a surprised facial expression) or flat (like in an angered expression).

Features F15 and F16 are the measures of the inner angles formed by the outter corner and center points of the eyebrows with the inner corners of the eyebrows. They have the same purpose of the features F13 and F14.

Some of the points used to calculate the features aren't directly output by the face tracker algorithm adopted in this work, and must be calculated before the features can be computed. These points are: UEl_{m3}, UEr_{m5}, EBl_{aux} and EBr_{aux}. The Equations 2, 3, 4 and 5 describe how each of these points are obtained. EBl_{aux} and EBr_{aux} are not facial landmarks, but auxiliar points used in conjunction with the landmarks to calculate some of the chosen features.

$$UEl_{m3} = \left(\frac{Ell_{Mx} + Elr_{Mx}}{2}, \frac{Ell_{My} + Elr_{My}}{2}\right) \tag{2}$$

$$UEr_{m5} = \left(\frac{Erl_{Mx} + Err_{Mx}}{2}, \frac{Erl_{My} + Err_{My}}{2}\right) \tag{3}$$

$$\begin{aligned} UEr_{m5} = (&Elr_{Mx} - Erl_{Mx} + EBlr_{Mx}, \\ &Elr_{My} - Erl_{My} + EBlr_{My}) \end{aligned} \tag{4}$$

$$\begin{aligned} UEr_{m5} = (&Erl_{Mx} - Elr_{Mx} + EBrl_{Mx}, \\ &Erl_{My} - Elr_{My} + EBrl_{My}) \end{aligned} \tag{5}$$

C. Databases

Once the feature set is chosen, the next step is to choose one or more databases to extract these features from. These databases should contain samples of all of the concepts the machine learning algorithm is expected to learn.

In the present work, both Cohn-Kanade Plus[30] and MMI Facial Expression [31] Databases are used to train the instantaneous facial emotion classifier model.

The Cohn-Kanade Plus (or CK+) Database comprises of 486 sets of pictures from 97 posers. Each set contains a sequence of pictures depicting a person acting the onset of a particular target emotion and each sequence is labeled as a sample of that particular represented target emotion. All of the sets start with a neutral expression and evolve into a particular target emotion expression.

The CK+ Database contains, but is not limited to, sequences of all of the studied basic emotions, that is: Happiness, Sadness, Anger and Fear; but doesn't contain sets labeled as Neutral. For the purpose of this work, for each selected set, the first picture of the sequence is taken as a Neutral sample and the last picture of the sequence is taken as a sample of the sequence's target emotion. To avoid one emotion being predominant over the others in the training set, which could degrade the quality of the training process, the limit of samples for each target emotion is set to be the number of samples available for the scarcer target emotion. After the features are extracted from the chosen sets, 129 samples are generated by this process.

The MMI Facial Expression Database comprises of over 2900 videos and images of 75 posers. Only part of these videos are labeled as samples of basic emotion, so just a subset of the database is effectively utilized in this work. The selected videos show humans acting a full emotional cycle of a particular target emotion, that is, all of the three phases of the emotional display are represented: onset, apex and offset. All of the selected videos start with a Neutral face expression, which progresses to a target emotion expression and then regresses back to the Neutral display.

Similarly to the CK+ Database, the MMI Facial Expression Database contains, but is not limited to, videos of all of the studied basic emotions, but doesn't contain samples of Neutral displays. Since the videos aren't labeled at a frame-level and considering there is no preliminary indication of which of the frames represent the emotion's apex, one must first manually annotate the frames' target emotions before they can extract the features from them.

That said, all of the 74 videos chosen from this database were annotated in the following manner: the authors would watch the videos and pinpoint four instants of interest. The first instant (referred to as t_1 from here forth) represents the start of the emotional onset in the video; the second instant (t_2) represents the emotional onset's ending and the beginning of the apex; the third instant (t_3) represents the apex's ending and the beginning of the emotional offset; finally, the fourth instant (t_4) represents the emotional offset's ending.

With these instants annotated, a frame-level categorization of the videos is created: the frames before t_1 and after t_4 (inclusive) are classified as Neutral samples; the frames between t_2 and t_3 (inclusive) are classified as that video's target emotion samples; finally, the frames between t_1 and t_2 and t_3 and t_4 are classified partially as Neutral samples and partially as that video's target emotion samples.

However, not all of the generated samples were used to train the classifier. The first and the last frames of each video were chosen to compose the Neutral set of the database; also, windows of size $n = 10 frames$ were built around the center of the apex region (that is, around the middle frame between t_2 and t_3) in each video, and all of the frames within these windows were taken as samples of that video's target emotion. The value of n was chosen empirically, and aimed to stablish a balance between the quantity of Neutral samples and the quantity of the other four emotions' samples. Also, care was taken so the created windows would never exceed their boundaries, that is, a window would never start at an instant before t_2 nor would it end after t_3.

After the features are extracted from the chosen pictures, 809 samples are generated by the described process.

It's worth saying both of the adopted databases contain videos and images of faces in profile and in other non-frontal orientations. However, different head orientations may cause the selected features to vary considerably for samples of the same target emotion. This could hamper the classifier's learning process and, for that reason, only videos and images containing emotional displays in frontal-oriented faces are used to train the classifier.

Finally, one should take note that all the sample images contained in these databases were acted, and not naturally elicited.

D. Instantaneous Facial Emotion Recognizer

The instantaneous facial emotion recognizer is a machine learning algorithm trained over the training set extracted through the previously described procedure.

The Random Forest learning algorithm is adopted in this work, as it was shown to have good accuracy on the work of Loconsole [28] when compared to other algorithms. The learner's accuracy and other statistics of interest are presented further in Section IV

The information fed into the dynamic classifier, however, is not simply the category output by the instantaneous classifier for a given sample, but rather, a measure of confidence that the classifier has for that sample to belong to each of the considered classes. The confidence measure used was the normalization of the number of votes each class received by the weak learners. Suppose, as an example, that a particular sample is classified by a random forest containing 100 random trees, and that 70 trees vote for the sample to belong to the Happiness class and the rest of the trees vote for it to belong to the Neutral class; in that case, the confidence measure for the sample to belong to the Happiness class would be 70%, the confidence measure for the sample to belong to the Neutral class would be 30% and the confidence measure for the sample to belong to the other classes would be 0. So, given a sample S_1, the output of the instantaneous classifier that is fed into the dynamic model is a vector of the form $V_1 = (Pr_{1n}, Pr_{1h}, Pr_{1s}, Pr_{1a}, Pr_{1f})$, where $Pr_{1n}, Pr_{1h}, Pr_{1s}, Pr_{1a}$ and Pr_{1f} are the confidence levels for S_1 to belong to the Neutral, Happiness, Sadness, Anger and Fear categories, respectively.

E. Kalman Filter

After the instaneous facial emotion classifier is properly trained, its outputs can be fed into the dynamic classifier, which will output the model's final prediction for the samples. However, aiming to eliminate high frequency noises, these outputs are firstly processed by a Kalman filter before they are inserted into the dynamic model. This section describes this filter and highlights the advantages of its usage.

As a natural consequence of the use of video frames to analyze the facial features of a person, different sources of noise can affect the classification algorithm.

It is assumed that the emotions are represented by the data initially fed in the training phase, which are gathered under controlled conditions; thus, effects such as face deformation resulting from speech, light source variations and unexpected face motions should be minimized. Furthermore, the objective of the model is to enhance the presentation of the slow and continuous emotions in spite of the instantaneous ones, sp a low pass filter should be used.

Kalman filtering is the solution proposed to this model, being a filter that has a good performance on linear systems with zero mean Gaussian noise on both the model and in the process of data acquisition. The empirical evidence presented in [9] supports this choice.

Being x_s the state variable of a linear system and y, the output of the filter for a single emotion, the filtered signal related to one of the emotions being analyzed, 6 and 7 describe the Kalman filter.

$$\dot{x_s} = x_s \tag{6}$$

$$\dot{F_a} = y = \frac{Kx_s}{\tau} \tag{7}$$

In the above equations, K is the filter's gain and τ is the time constant. There are two steps for the filtering, the first being the prediction step and the second the update step. The update is only run when new information from the sensors – in this case the output of the instantaneous emotion analyzer – is available. If the delay between data acquisitions is higher than the delay between filter steps calculations there will be some steps in which only the prediction steps will be run.

The prediction step is described by 8 and 9, where $x_{s,t}$ is the current state $x_{s,t-1}$ is the previous state, w is the noise covariance, p the covariance of the state variable on the t state. Note that the update step always assumes that the state variable has not changed, only the covariance of the system.

$$x_{s,t} = x_{s,t-1} \tag{8}$$

$$p = p + w/\tau^2 \tag{9}$$

The update step is described by equations 10, 11 and 12, where m is the residual covariance, v is the covariance of the observation noise and r_t and y_t are the filter input and output at instant t. This input corresponds to the output of the instantaneous emotion classifier. The state variable now has its value updated and, consequently, the output of the Kalman filter has its value proportionally changed.

$$m = \frac{\frac{pK}{\tau}}{p\left(\frac{K}{\tau}\right)^2 + v} \tag{10}$$

$$x_{s,t} = x_{s,t} + m(r_t - y_t) \tag{11}$$

$$p = (1 - \frac{mk}{\tau})p \tag{12}$$

Note that these equations describe the filtering process for a single class (that is, the filtering of outputs of a particular emotion). The full model is represented by applying these equations for each emotion separately.

However, neither w nor v are known, and have to be estimated by an optimization algorithm, which is described in Section III-G.

F. Dynamic Model

After the instantaneous output is filtered, it is ready to be fed into the dynamic model.

The dynamic model proposed here does not aim to describe rapid emotional variations a person may be subject to, but rather, it tries to describe more lasting emotional states. Suppose, for illustration purposes, that a man is talking to a dear friend of him that he has not seen for a while. One may expect the overall conversation to elicit a pleasant emotion. However, during this conversarion, he happens to see a person throwing trash in the street; it infuriates him for a while, but he rapidly get back to talking to his friend and forgets the sight that angered him. If pictures of his face were fed into the proposed model during this entire event, one should expect the model to detect the overall pleasant emotion of the conversation (that is, if it was pleasant enough so that his facial expression indicated so); however, his temporary enragement should not modify the output of the model.

The dynamic model is based on the work of [9], and utilizes the concept of Dynamic Emotional Surfaces (DESs).

As name indicates, DESs are surfaces that aim to describe the dynamics of transitions between different emotional states. In this work, a planar surface is adopted, and it is partitioned in four quadrants, one for each of the considered basic emotions: Happiness, Sadness, Anger and Surprise. Centered in the intersection of the four areas, there is the Neutral area, which represents the absence of emotions. Figure 4 illustrates the model's DES.

Located on the $+45°$ and $-45°$ diagonals of this plane, there are four Emotional Attractors (EAs), one for each of

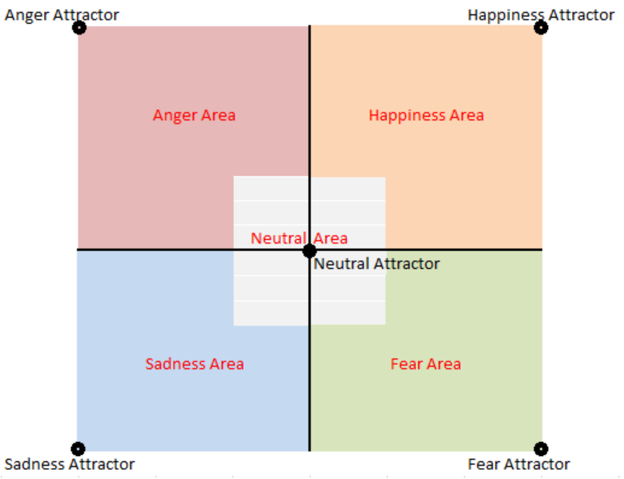

Fig. 4: A representation of the planar DES used in this work.

the considered non-Neutral emotions, and each located in its corresponding quadrant. Refer to Figure 4. The Happiness, Fear, Sadness and Anger attractors are located on the points $PA_{Happiness}$, PA_{Fear}, $PA_{Sadness}$ and PA_{Anger}, respectively, and the Neutral attractor is located on the point $PA_{Neutral}$.

Let to slide upon the plane, there are Emotional Particles (EPs), one for each analyzed subject. The location of a particle in a given instant indicates the model's output emotion for that instant, according to the equation below, where $P = (P_x, P_y)$ is an EP's position and $f(P)$ is the model's output in the considered instant.

$$f(P) = \begin{cases} \text{Happiness}, & \text{if } P_x > K_{nr} \text{ and } P_y > K_{nr} \\ \text{Sadness}, & \text{if } P_x < -K_{nr} \text{ and } P_y < -K_{nr} \\ \text{Anger}, & \text{if } P_x < -K_{nr} \text{ and } P_y > K_{nr} \\ \text{Fear}, & \text{if } P_x > K_{nr} \text{ and } P_y < -K_{nr} \end{cases} \tag{13}$$

In the equation 13, K_{nr} is a constant that determines the width and height of the Neutral area.

The EAs are responsible for pulling EPs towards them. The stronger the confidence level the Kalman filter outputs for a given emotion, the stronger the pull velocity for that emotion's attractor will be. If at the instant $\bar{t}$ Kalman filter outputs a confidence level of $Pr_E(\bar{t})$ for emotion E, then E's attractor velocity, $VA_E(\bar{t})$, is given by Equation 14.

$$VA_E(\bar{t}) = K_{avm} Pr_E(\bar{t}) \tag{14}$$

The parameter K_{avm} is the attractors velocity modulator parameter, which value, like the Kalman filter parameters w and v, is also found via an optimization algorithm.

The dynamics for EPs are described by equations 15 and 16, where $P(t)$ and $V(t)$ are particles' position and velocity, $Pr_E(t)$ is the confidence measure for emotion E and $VA_E(t)$ is the attractor's E pull velocity, all at instant t.

$$P(t) = P(t-1) + V(t) \tag{15}$$

$$V(t) = \begin{cases} VA_{Neutral}, & \text{if } max(Pr_E(t)) = Pr_{Neutral}(t) \\ \sum_{E=\bar{e}} VA_E, & \text{if not} \end{cases} \tag{16}$$

where $\bar{e}$ is the subset $\{Happiness, Sadness, Anger, Fear\}$. Also, the position of the particle is never let to exceed the rectangle delimited by the four non-Neutral EAs.

The noise smoothing introduced by the Kalman filter and the intrinsic inertia presented by the proposed model make so that natural facial noises - like the mouth movements caused by laughter or speech - should have its influence on the predictions diminished, when in comparison to the instantaneous classifier.

G. Parameters Optimization

As the previous sections explained, some of the model parameters can not be known *a priori*, and are better defined via an optimization process. These parameters are: Kalman filter's noise covariance (w), Kalman filter's covariance of the observation noise(v) and DES's attractors velocity modulator (K_{avm}).

The optimization process here adopted is based on the simulated annealing algorithm, and can be described by the pseudo-code presented below.

In the above pseudo-code, T_0, T_{room} and T_{curr} are the initial, room and current temperatures of the optimizer, in that order; p_curr, p_{la} and p_{sol} are the current iteration's parameters, the last accepted solution's parameters and the final solution parameters, respectivelty; e_0, e_{curr}, e_{la} and e_{sol} are the initial energy, the current iteration's energy, the last accepted solution's energy and the final solution's energy, in that order; dr is the temperature decay rate and Pr_{acc} is the probability that a solution will be accepted by the algorithm.

Note that an iteration's energy, e_{curr} is obtained by the function $calculateEnergy(p_{curr}, dataset)$, which considers both the current value of the parameters being optimized and a dataset chosen for the optimization. The MMI Facial Emotion Database's previously selected 74 videos were used to extract the energy measure; however, this time they were considered in their full-length. The adopted energy measure is the number of frames the model misclassified in the iteration.

A proposed solution is always accepted if it causes the system's energy to decrease in comparison to the last accepted solution's energy. However, even if a solution causes the energy to increase, it has a chance of being accepted that is proportional to the iteration's current temperature and inversely proportional to the energy increase it causes. This measure helps the optimizer to avoid getting stuck in local minima.

If a solution is accepted, its parameters and energy are stored to serve as comparison data for the next iteration. However, a solution is only stored as a final solution if its energy is smaller than the last accepted final solution.

```
// Initializations:
$T_0 = 200°C$;
$T_{room} = 20°C$;
$T_{curr} = T_0$;
$p_{curr} = randomizeParameters()$;
$p_{la} = p_{curr}$;
$p_{sol} = p_c$;
$e_0 = +\infty$;
$e_{curr} = e_0$;
$e_{la} = e_0$;
$e_{sol} = e_0$;
$dr = 0.99995$;
// Iterations:
while ($T_{curr} > T_{room}$) do
    $e_c = calculateEnergy(p_{curr}, dataset)$;
    if ($e_{curr} < e_{la}$) then
        $Pr_{acc} = 1$;
    else
        $Pr_{acc} = e^{(e_{la} - e_{curr})/T_{curr}}$;
    end
    if ($Rnd(0,1) > Pr_{acc}$) then
        $p_{la} = p_{curr}$;
        $e_{la} = e_{curr}$;
        if ($e_{la} > e_{sol}$) then
            $p_{sol} = p_{la}$;
            $e_{sol} = e_{la}$;
        else
            $p_{curr} = p_{la}$;
        end
    end
    $p_{curr} = moveAround(T_{curr})$;
    $T_{curr} = T_{curr} \times dr$;
end
```

At the end of every iteration, the parameters are varied through the function $moveAround(T_{curr})$, which takes into consideration the iteration's temperature - the higher the temperature, the more the parameters are allowed to variate -, and the optimizer temperature is made to decay by a constant rate dr.

IV. Tests and Results

This section presents the results of the tests realized on the model. These tests are presented separately for the instantaneous facial emotion classifier, for the parameters optimization algorithm and for the dynamic facial emotion classifier.

A. *Tests on the Instantaneous Facial Emotion Classifier*

Tests were made to measure the quality of the instantaneous classifier. Since a poorly trained classifier may compromise the overall performance of the model, the quality of its outputs should be analyzed with caution.

The random forest learning algorithm discards the need for procedures like cross-validation, bootstrap or separate test sets for estimating the classifier's accuracy. During the training of each of the weak learners (that is, of each tree of the forest), an out-of-bag set (or "oob set", containing roughly 1/3 of the complete training set) is created for that learner. The oob set is used to validate the accuracy of that particular tree. After all the trees have finished training, the following procedure is used to calculate an estimation of the accuracy of the learner for samples stranger to the training set:

1) Each sample contained in the complete training set is considered separately;
2) All trees that contain a particular sample in their oob sets are used to classify that sample, and a vote counting is used to decide to what class it belongs to. The procedure is repeated for all samples of the complete training set;
3) The random forest's accuracy is given by the number of samples of the set classified correctly divided by the number of samples classified incorrectly by that process....

Through the described procedure, an accuracy estimate of approximately 90% was obtained.

The analysis of the learner's confusion matrix allows one to observe how its predictions are distributed amongst the different classes. Table II presents the confusion matrix for the trained random forest.

Table II: The confusion matrix for the trained random forest

	Neutral	Anger	Fear	Happiness	Sadness
Neu.	201	5	4	8	20
Ang.	0	168	1	0	5
Fea.	0	0	168	0	1
Hap.	0	0	2	189	0
Sad.	0	1	0	0	113

One can notice that the Neutral class is the one with more misclassified samples, even if considering relative numbers. Also, more than half of the misclassified Neutral samples are categorized as Sadness samples, which suggests that the boundaries between these two classes is the less obvious for the classifier, at least on the considered dataset.

B. *Tests on the Parameters Optimization Algorithm*

Tests were made with values of K_{nr} (which determines the height and width of the neutral area of the plane) varying from 1 to 5, with unitary increments. Also, for all the tests, the attractors were positioned on the points $PA_{Happiness} = (10, 10)$, $PA_{Fear} = (10, -10)$, $PA_{Sadness} = (-10, -10)$, $PA_{Anger} = (-10, 10)$ and $PA_{Neutral} = (0, 0)$. Figure 5 shows the model's accuracy history for the best optimization achieved - that is, for the optimization that reached the lowest energy on the used dataset.

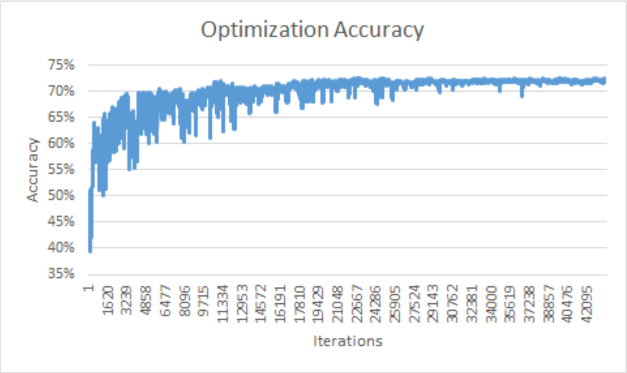

Fig. 5: The accuracy curve for the best optimization case.

It is possible to see an overall increase of the optimization accuracy as the iterations progress; also, the accuracy curve seems to converge to a value of about 72% by the end of the process. The accuracy achieved with the instantaneous model for the same dataset is of 64%. This fact suggests that the use of the dynamic model was beneficial even for a dataset with videos that do not contain too many facial noises caused by factors like laughter or speech.

The best accuracy was reached for a value 1 of K_{nr}.

C. Tests on the Dynamic Facial Emotion Classifier

To test the developed dynamic classifier, a test was run on the video "S43_an_2" of the eNTERFACE'05 Database [32], the same analyzed in the work of [9].

This video depicts the face of an angered person as she irritatedly proclaims a certain sentence. The presence of facial noises in the video is relevant for the experiment, as it allows for the analysis of how well the dynamic model is able to deal with such noises. Also, this is the first experiment that utilizes a video entirely stranger to the datasets used for training the instantaneous classifier and for the optimization process.

Because the video "S43_an_2" is simply classified as an Anger video, and since there is no information about whether any other emotional displays are considered to be present in it, all of its frames were considered as Anger samples and fed into the model.

Figures 6 and 7 present the dynamic model output and the instantaneous classifier output for each frame of the video, respectively. The accuracy achieved with the dynamic model was of 89%, while the accuracy achieved with the instantaneous classifier was of 64%. This result suggests that the dynamic model successfuly dealt with a considerable portion of the facial noises presented in the video. Note that not only the dynamic model achieved a higher accuracy on the video, but its outputs seem to be more reliable. With exception of the last frame, all frames in the video were classified as Anger or Neutral frames by the dynamic model, and there are less variations between different emotional states; the classifications attributed by the instantaneous model, however, flicker more rapidly and between a larger number of emotional states. One could argue that the result achieved by the dynamic model is more useful than the one achieved by the instantaneous classifier if it was to be used to control an automated system like a social robot - maybe the social robot would not be able to react as well to a flickering input as it would react to a more stable one.

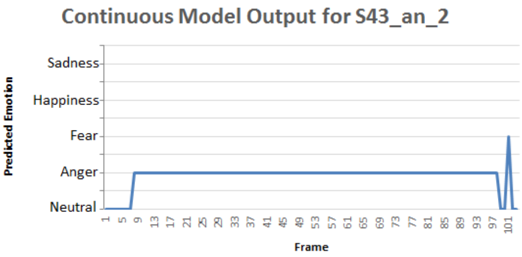

Fig. 6: The output of the continuous model for the video "$S43\_an\_2$".

Finally, Figure 8 presents the trajectory on which the EP traveled throughout the video. Note that the particle rapidly progresses to the Anger area, where it remains until the latter parts of the video, regressing back to the neutral area and then to the Fear area by the end of the video. The transition to the Fear area is probably due to the considerably large number of Fear predictions outputted by the instantaneous classifier in the latter parts of the video.

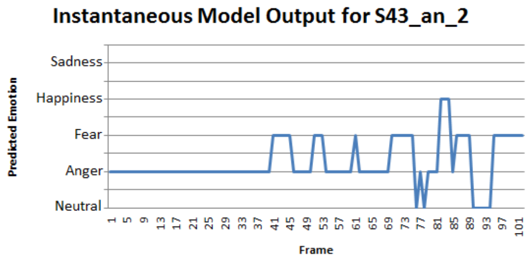

Fig. 7: The output of the instantaneous classifier for the video "$S43\_an\_2$".

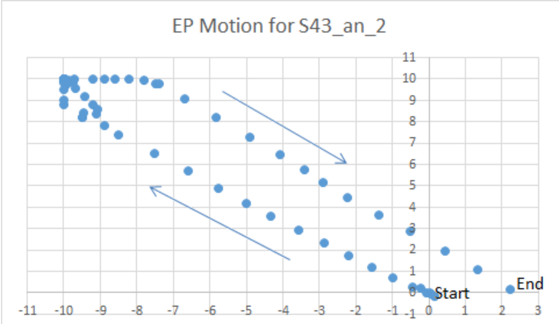

Fig. 8: Motion of the EP throughout the video "$S43\_an\_2$".

V. Conclusions

In the present work, an innovative dynamic emotion recognition model was presented. This model comprises of the conjugation of a machine learning algorithm, a Kalman filter and an original dynamic model that aims to describe durable emotional states and to minimize facial noises like deformations caused by laughter and speech. A simulated annealing algorithm was utilized to optimize the model's parameters.

The model has shown good performance when compared to the instantaneous emotion classifier trained in the present work: while the former achieved an accuracy rate of 72% over the chosen dataset, the latter presented an accuracy rate of just 64%, on the same dataset.

When tests on a sample stranger to the datasets utilized to train the instantaneous classifier and to optimize the model's parameters, the dynamic model once again outmatched the instantaneous model: not only it achieved a higher accuracy rate (89% against 64%), but it also provided a much more stable output.

As target objectives for future works, the following tasks are proposed:

1) Execute more tests on the dynamic model, in order to better analyze its accuracy and the way it describes the progression of emotional expressions in faces;

2) Utilize larger datasets to train the instantaneous model and to optimize the dynamic model;
3) Utilize datasets that contain faces deformed by natural facial noises, like laugther or speech, for the training and optimization of the model;
4) Study possible changes the proposed planar DES may need to better describe the way emotions manifest themselves in human faces;
5) Increase the number of considered emotions and study how the DES should be changed to accommodate this change.

REFERENCES

[1] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power, Eds. Chichester, UK: John Wiley Sons, Ltd, 2005, ch. 3, p. 45–60.

[2] R. W. Picard, "Affective computing," MIT Media Lab, Perceptual Computing, Cambridge, MA, Tech. Rep. 295, 1995.

[3] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[4] P. Ekman, "A linguagem das emoções," *São Paulo: Lua de Papel*, 2011.

[5] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[6] R. A. M. Gonçalves, "Um modelo matemático para inferência computacional de estado emocional a partir de detectores de expressões faciais," M. Eng. thesis, Universidade de São Paulo, 2012.

[7] Z. Zeng, M. Pantic, G. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[8] M. Zarkowski, "Identification-driven emotion recognition system for a social robot," in *Methods and Models in Automation and Robotics (MMAR), 2013 18th International Conference on*. IEEE, 2013, pp. 138–143.

[9] R. A. M. Gonçalves, D. R. Cueva, M. R. Pereira-Barretto, and F. G. Cozman, "A model for inference of emotional state based on facial expressions," *Journal of the Brazilian Computer Society*, vol. 19, no. 1, pp. 3–13, 2013.

[10] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide, "Evaluation and discussion of multi-modal emotion recognition," in *Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on*, vol. 1. IEEE, 2009, pp. 598–602.

[11] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, 1989.

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[13] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural networks*, vol. 18, no. 4, pp. 317–352, 2005.

[14] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[15] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[16] B. Tinen, "Sistema de identificação de emoções por expressões faciais com operação ao vivo," Eng. thesis, Universidade de São Paulo, 2014.

[17] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movements," *Consulting Psychologist*, vol. 2, 1978.

[18] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 149–149.

[19] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[20] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[21] G. Donato, M. S. B. J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 974–989, 1999.

[22] S.-S. Liu, Y.-T. Tian, and D. Li, "New research advances of facial expression recognition," in *Machine Learning and Cybernetics, 2009 International Conference on*, vol. 2. IEEE, 2009, pp. 1150–1155.

[23] J. Hamm, C. G. K. R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237–256, 2011.

[24] T. F. Cootes, J. C. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[25] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[26] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[27] J. A. Russell, "Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

[28] C. Loconsole, C. R. Miranda, G. Augusto, A. Frisoli, and V. C. Orvalho, "Real-time emotion recognition-novel method for geometrical facial features extraction." in *VISAPP (1)*, 2014, pp. 378–385.

[29] A. Asthana, S. Zafeiriou, S. Cheng, and M. .Pantic, "Incremental face alignment in the wild," *Science*, 2014.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[31] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005.

[32] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.