
Article:
Garcia, H.F., Alvarez Lopez, M.A. and Orozco, A.A. (2017) Dynamic Facial Landmarking Selection for Emotion Recognition using Gaussian Processes. Journal on Multimodal User Interfaces, 11 (4). pp. 327-340. ISSN 1783-8738

https://doi.org/10.1007/s12193-017-0256-9

Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

Dynamic Facial Landmarking Selection for Emotion Recognition using Gaussian Processes

Received: date / Accepted: date

Abstract Facial features are the basis for the emotion recognition process and are widely used in affective computing systems. This emotional process is produced by a dynamic change in the physiological signals and the visual answers related to the facial expressions. An important factor in this process, relies on the shape information of a facial expression, represented as dynamically changing facial landmarks. In this paper we present a framework for dynamic facial landmarking selection based on facial expression analysis using Gaussian Processes. We perform facial features tracking, based on Active Appearance Models for facial landmarking detection, and then use Gaussian process ranking over the dynamic emotional sequences with the aim to establish which landmarks are more relevant for emotional multivariate time-series recognition. The experimental results show that Gaussian Processes can effectively fit to an emotional time-series and the ranking process with log-likelihoods finds the best landmarks (mouth and eyebrows regions) that represent a given facial expression sequence. Finally, we use the best ranked landmarks in emotion recognition tasks obtaining accurate performances for acted and spontaneous scenarios of emotional datasets.

Keywords Facial landmark · Dynamic emotion · Statistical models · Gaussian Processes · Gaussian Process Ranking

1 Introduction

Facial landmarking analysis plays an important role in many applications derived from face processing operations, including emotion recognition, facial animation, and biometric recognition [5]. Analyzing this information is particularly useful in the emotion recognition field, because from the landmarks analysis (eg. eye corners, eyebrows, mouth corner etc.), we can describe a given facial expression. Since facial expressions are characterized by changes of facial muscles, we need to capture the dynamic changes on facial features. Moreover, these temporal facial changes are mainly represented by the facial shape information derived from the facial landmarks [12].

Despite that facial feature detection methods to locate facial landmarks have been widely investigated in the state of the art, the computer vision problem has proven extremely challenging for emotion recognition systems derived from facial landmarking analysis [31]. In order to define which facial features are more relevant to recognize an emotion type, most of the works for emotion recognition are based on Ekman's study [9]. This study, proposes a comprehensive and anatomically based system called Facial Action Coding System (FACS), which is used to measure all visually discernible facial movements in terms of atomic facial actions called Action Units (AUs). As AUs are independent of interpretation, they can be used for any high-level decision-making process, including the recognition of basic emotions according to Emotional FACS (EMFACS), the recognition of various affective states according to the FACS Affect Interpretation Database (FACSAID) [10], and the recognition of other complex psychological states such as depression or pain [9].

Since AUs are suitable to use in studies on facial behavior (as the thousands of anatomically possible facial expressions can be described as combinations of 27 basic AUs and a number of AU descriptors [11]), it is not surprising that an increasing number of studies on human spontaneous facial behavior are based on

automatic AU recognition [6, 32]. Furthermore, facial landmarks are used to compute these AUs in order to perform the facial expression analysis [7, 38]. The typical facial features used to perform the facial expression analysis are either morphological features such as shapes of the facial regions (eyes, nose, facial contour, mouth, etc.), and the location of facial salient points (corner of the eyebrows, mouth, chin tip, etc.) [27]. However, these methods do not address which points are most relevant in the analysis of a sequence of facial expressions.

Pantic and Valstar have reported different studies for facial action units recognition and their temporal segments [34]. In all their studies only a specific set of facial landmarks (corners of the eyebrows, eyes, mouth and nose tip) are used to compute the action units [25]. Since facial expressions are different for every subject, due to characteristic features of each person in the manifestation of a particular emotion, it would be important to propose a methodology to study a wide range of facial points and their temporal dynamics, in order to recognize a much larger range of expressions (apart from the prototypic ones i.e spontaneous facial expressions) [13]. Moreover, most of the works in the emotion recognition field are based on using the entire facial shape model (or use some salient points such as nose tip, corners of eyes, mouth and eyebrows) [33]. Since spontaneous emotional behaviors vary depending from how people perceive their environment [30], it is required to analyze which specific facial landmarks brings more relevant information in an emotional sequence. The main research topic in this paper, is to model the temporal activity of each facial landmark, and to rank those facial features that describe an emotional process [13].

Furthermore, it is worth mentioning that there is a number of works in the state-of-the-art, in which dynamic analysis is used for emotion recognition [41]. Here, most of these works used physiological signals such as electroencephalogram, electromyogram, respiration and heart rate in order to perform the recognition [28, 24, 37]. However, works that uses facial expressions features for dynamical analysis, are based only in computing Action Units as features, but discards modeling the temporal changes of the facial landmarks [28].

Due to the need of modeling the dynamics of facial features in an emotional sequence, we use supervised learning for regression tasks. Commonly parametric models have been used for this purpose. These have a potential advantage of ease of interpretability, but for complex data sets, simple parametric models may lack generalized performance. Gaussian processes (GP) [29, 21], offer a robust way to implement approaches to quantify the dynamic facial features embedded in a facial expression time-series, and thus allow us to rank the set of facial features that best depicts a dynamical facial expression.

In a dynamic facial expression framework, a Gaussian process coupled with the squared-exponential covariance function or radial basis function used in regression tasks, can efficiently perform dynamic facial feature selection in emotional time-series [30]. This property outfit the GP with a wide degree of flexibility in capturing the dynamic landmark variations of facial features. Moreover, this property makes the GP an attractive novel tool for affective computing applications [16].

In this work, we develop a novel technique for facial landmarking selection by analyzing the dynamical visual answers of the facial expressions (specially those in regions of FACS [26]) using Gaussian processes. Those features are detected by using statistical models as Active Appearance Models (AAM) proposed in [22], which from the prior object knowledge (face to be analyzed), allow us to estimate the object shape with high accuracy. From the facial features detected, it is possible to estimate which landmarks are more relevant in a specific dynamic facial expression. The proposed method employs a Gaussian process for regression over the dynamical facial features with the aim to identify which landmarks are more relevant in the dynamical emotional process. Facial features are ranked according to the signal to noise ratio (SNR), which is captured by fitting a Gaussian process. In addition, a statistical analysis on multiple datasets is performed to verify the generalizability of the proposed method. Finally, we use the best ranked facial landmarks for every emotional-time series, and then perform a dynamic classification task based Hidden Markov Models (HMM) for validation purposes. The main contribution of this work is the development of a methodology that is able to rank the facial landmarks that are more relevant in a dynamic emotion sequence when it comes to emotion recognition.

The paper includes the following sections. Section 2.2 presents the facial feature extraction model used in this work. Section 2.2 presents our facial landmarking selection method. Sections 2.4 and 3 discuss the experimental setup and results respectively. The paper concludes in Section 4, with a summary and discussion for future research.

Table 1 Facial expression databases used in this work.

Name	No. of sequences	Expression/ Pose changes/ illumination	Color/ Gray	Resolution	Number of subjects	Year
Cohn-Kanade database [19]	327	7, No, No	Gray	640x490	123	2003
FEED database [35]	399	7, No, Yes	Color	320x240	18	2006
Oulu-CASIA database [39]	480	6, No, Yes	Color	320x240	80	2011
RML emotion database database [36]	200	6, Yes, Yes	Color	640x480	7	2008

2 Materials and Methods

2.1 Database

In this work four databases were used (See Table. 1 for a description). The first database is the Cohn-Kanade AU-Coded Facial Expression Database. It was developed for research in automatic facial image analysis and synthesis for perceptual studies [19]. It includes both posed and non-posed (spontaneous) expressions and additional types of meta-data (files with images, facial landmarks, action units and emotional labels). The target expression for each sequence is fully FACS coded. In addition, validated emotion labels have been added to the meta-data. Thus, sequences may be analyzed for both action units and prototypic emotions [9] (See Fig. 1). The second database is the FEED Database with Facial Expressions and Emotions from the Technical University of Munich containing face images showing a number of subjects performing the six different basic emotions defined by Eckman & Friesen [35] (See Fig. 2). The database has been generated as part of the European Union project FG-NET (Face and Gesture Recognition Research Network). The database contains material gathered from 18 different individuals and each individual performed all six desired actions three times. Additionally three sequences with no expressions at all are recorded. Altogether, this gives an amount of 399 sequences. The third database is the Oulu-CASIA facial expression database [39]. This database was developed by the Machine Vision Group of the University of Oulu, which consists of six typical expressions (surprise, happiness, sadness, anger, fear and disgust) from 80 people between 23 to 58 years old. Subjects were asked to make a facial expression according to an expression example shown in a given sequence (acted facial expression) (See Fig. 3).

Finally, the fourth database, is the RML emotion database, for which we used 60 (ten for each emotion) spontaneous audiovisual emotional expression samples that were collected at Ryerson Multimedia Lab at the Ryerson University (See Fig. 4). Six basic human emotions are expressed: Anger, Disgust, Fear, Happiness, Sadness, Surprise. The RML emotion database is suitable for audio-based, static image-based, and video-based 2D and 3D dynamic analysis and recognition [36].

Fig. 1 Description of an emotional expression for the Cohn Kanade database in which it can be seen the emotional process (acted) starting from neutral to peak expression (lower right).

Fig. 2 Description of an emotional sequence for the FEED database in which it can be seen the emotional process starting from neutral to peak expression.

Fig. 3 Description of an emotional sequence for the Oulu-CASIA facial expression database, in which the emotional process (acted) was recorded in a weak illumination scenario.

Fig. 4 Sample images of RML emotion database showing an spontaneous emotional sequence.

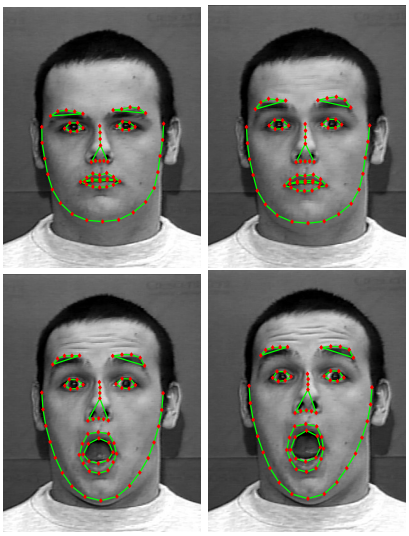


Fig. 5 Facial description for CK database using a set of 68 landmarks to describe the shape model.

2.2 Active Appearance Models for facial feature extraction

An Active Appearance Model (AAM) is built through a process of learning marked features for a class of objects. An AAM allows us to find the parameters of such a model which generates a synthetic image as close as possible to a particular target image, assuming a reasonable starting approximation [22].

2.2.1 Landmarking The Training Set

In order to build our facial feature extraction method (AAM), we select the facial images to incorporate into the training set. Here, an important task is to decide which facial images will be included in the training set. To this end, the desired variations of facial expressions must be considered (i.e. prototypic facial expressions and those that include particular facial gestures). Therefore, we labeled the facial expressions defined by Ekman's study in which the basic emotions are covered (i.n. happiness, anger, fear, disgust, sadness, contempt and surprise) [9]. However, due to the need of modeling spontaneous emotional behaviors, we landmark those facial expressions related to the RML database. Here, ten subjects were labeled in order to add these spontaneous facial expressions in the recognition process. To build the AAM model, we used 50 emotional sequences from the CK database, 50 sequences from the FEED-TUM and 50 sequences from the RML database. From these sequences, we model all those shape variations related with the prototypical emotions and the spontaneous ones (i.e. a given facial expression).

We use the parametrized face model used for the CK database [19]. Here, a set of 68 landmarks were labeled for each image in the dataset, in order to describe the facial shape (landmarks for describing eyes, nose, mouth and eyebrows regions). Figure 5, shows an example of the shape model used to depict a facial expression.

2.2.2 Facial landmark detection

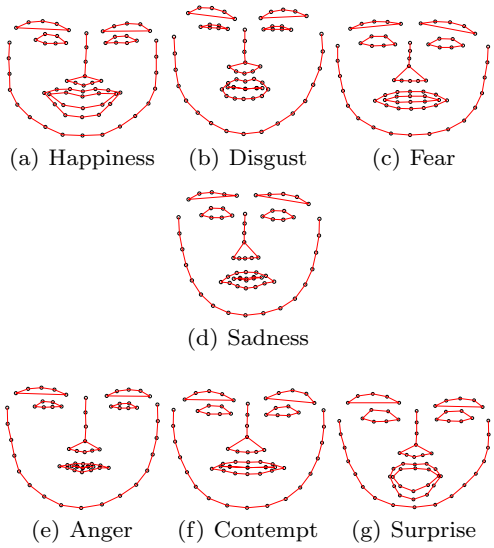
An AAM contains a statistical model of the shape and grey-level appearance of the object of interest which can generalize to almost any valid example (labeled facial expression). Matching to an image involves finding model parameters (shape and appearance descriptors) which minimize the difference between the image and a synthesized model example, projected into the image. The potentially large number of parameters makes this a difficult problem. We use the proposed facial landmarks detection method proposed by Edwards *et. al.* in [22]. An AAM algorithm is applied for building the shape and appearance models. Then, facial landmarking detection is performed by fitting the built model to a facial sequence. We use the Active Appearance Model Face Tracker library using OPENCV in C++, to perform the facial tracking. [1] Figure 6 shows an example about how landmarks are located in the shape model for every emotion prototype.

To perform an error analysis over the landmark detection process, we compute the average error of the distance between the manually labeled points $\mathbf{p}_i$ and points estimated by the model $\hat{\mathbf{p}}_i$ for all training and test images. Also, to perform a quantitative analysis of the accuracy in adjusting of the AAM, we calculate the relative error between the manually labeled facial landmarks, and the points estimated by the AAM model for the eyelids and mouth regions.

2.3 Gaussian Processes

A Gaussian Process (GP) is an infinite collection of scalar random variables indexed by an input space such that for any finite set of inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

[1] Active Appearance Model Face Tracker library is available in `https://code.google.com/p/aam-opencv/`.



(a) Happiness (b) Disgust (c) Fear

(d) Sadness

(e) Anger (f) Contempt (g) Surprise

Fig. 6 Facial expression samples for each emotion.

the random variables $\mathbf{f} \triangleq [f(\mathbf{x}_1), f(\mathbf{x}_2), \cdots, f(\mathbf{x}_n)]$ are distributed according to a multivariate Gaussian distribution $\mathbf{f}(\mathbf{X}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ [23]. A GP is completely specified by a mean function $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})]$ (usually defined as the zero function) and a covariance function given by

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^\top\right].$$

We use a squared exponential kernel (Radial Basis Function (RBF) kernel) given by

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(\mathbf{x} - \mathbf{x}')^2\right), \quad (1)$$

where σ_f^2 controls the variance of the functions and l^2 controls the *lengthscale* which specifies the distance beyond which any two inputs $(\mathbf{x}_i, \mathbf{x}_j)$ becomes uncorrelated. Besides, if there are n inputs, we can write equation (1) in a matrix form, where $\mathbf{K}(\mathbf{X}, \mathbf{X})$ denotes the $n \times n$ matrix of the covariance evaluated at all pairs of inputs.

Moreover, by making predictions using noisy observations[2], given by $y = f(\mathbf{x}) + \epsilon$, the prior on the noisy observations becomes

$$\mathrm{cov}(\mathbf{y}) = \mathbf{K}(\mathbf{X}, \mathbf{X}) + {\sigma_n}^2\mathbf{I}, \quad (2)$$

where $\mathbf{K}(\mathbf{X}, \mathbf{X})$ denotes the covariance matrix.

By using the multivariate Gaussian properties, it is possible to obtain a predictive distribution $\mathbf{f}_*$ for new inputs $\mathbf{x}_*$ [4]. The Gaussian process regression is given by

$$\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \mathbf{cov}(\mathbf{f}_*)), \quad (3)$$

where

$$\bar{\mathbf{f}_*} \triangleq \mathbb{E}[\mathbf{f}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_*] = \mathbf{K}(\mathbf{X}_*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + {\sigma_n}^2\mathbf{I}]^{-1}\mathbf{y},$$

and

$$\mathrm{cov}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})\left[\mathbf{K}(\mathbf{X}, \mathbf{X}) + {\sigma_n}^2\mathbf{I}\right]\mathbf{K}(\mathbf{X}, \mathbf{X}_*).$$

To estimate the kernel parameters, we maximize the marginal likelihood which is faster than using exhaustive search over a discrete grid of values, with validation loss as an objective [23], [2]. Here, the marginal likelihood refers to marginalize the function values $\mathbf{f}$ [4].

Due to the fact that the prior for a Gaussian process is Gaussian, $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, and the likelihood is a factorized Gaussian $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, {\sigma_n}^2\mathbf{I})$ the log marginal likelihood is given by

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top\left(\mathbf{K} + {\sigma_n}^2\mathbf{I}\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|\mathbf{K} + {\sigma_n}^2\mathbf{I}\right| - \frac{n}{2}\log 2\pi. \quad (4)$$

From the equation (4), we can learn the hyperparameters, $\boldsymbol{\theta}$, from the data by optimizing the log-marginal likelihood function of the GP [4].

2.3.1 Ranking with likelihood-ratios

In this paper, we used the approach presented in [14], to estimate continuous trajectories of gene expression time-series through Gaussian process (GP) regression. Here, the differential expression of each profile were ranked via a log-ratio of marginal likelihoods. This approach was also used by [1], for selecting meaningful outputs from a motion capture dataset. To this end, we compute the Bayes factor with a log-ratio of marginal likelihoods (LML)[3], this factor is given by

$$\ln\left(\frac{p(\mathbf{y}|\mathbf{x}, \theta_2)}{p(\mathbf{y}|\mathbf{x}, \theta_1)}\right), \quad (5)$$

with each LML being a function of different instantiations of θ.

In order to rank with likelihood-ratios, we first set two different hypotheses[4] [14]. First, $\mathcal{H}_1$, represents how

[2] We assume additive independent identically distributed Gaussian noise ϵ with variance σ_n^2, given by $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$

[3] We use this Bayes factor instead of integrating out the hyperparameters θ (compute a posterior over hyperparameters). See equation (4).

[4] $\mathcal{H}_1$ and $\mathcal{H}_2$ represent two different configuration of the hyperparameters $\boldsymbol{\theta}$

the facial expression time-series (for each landmark) has a significant underlying signal (i.e. facial landmark that shows relevant changes in an emotional sequence). Second, $\mathcal{H}_2$ represents the fact that there is no underlying signal in the facial expression, and the observed facial sequence for a given emotion is just the effect of random noise.

From these hypotheses, we relate $\mathcal{H}_2$ with the hyperparameter $\theta_1 = (\infty, 0; \text{var}(\mathbf{y}))^\top$ to model a function constant in time ($l^2 \to \infty$), with no underlying signal ($\sigma_f^2 = 0$). This process generates a facial landmark time-series, with a variance that can be solely explained by noise ($\sigma_n^2 = \text{var}(\mathbf{y})$ *white kernel* in the GP). Finally on $\mathcal{H}_1$, the hyperparameters θ_2 are set to model a facial landmark sequence that fits an emotional process. Here, we use a distinct signal variance that solely explain the observed facial sequence variance ($\sigma_f^2 = \text{var}(\mathbf{y})$) and with no noise ($\sigma_n^2 = 0$).

2.4 Procedure

After estimating facial landmarks belonging to an emotional sequence[5], we proceed to train Gaussian processes to measure dynamically which of these landmarks are more relevant in a given emotional sequence. We use the GPmat toolbox for Gaussian Process Ranking developed in [14]. [6] The following steps depicts the proposed method for facial landmarking selection using Gaussian processes.

1. For every landmark in an emotional sequence.
   (a) Train a GP following the two hypothesis depicted in section 2.3.1
   (b) Compute the log-ratio of marginal likelihood (see equation (5)). Here, The facial landmark ranking is based on how likely $\mathcal{H}_1$ in comparison to $\mathcal{H}_2$, given a facial expression sequence from the signal-to-noise ratio SNR $= \sigma_f^2/\sigma_n^2$.
2. Rank the log-like ratios for every landmark in all dynamical emotional expressions.

Figure 7 shows the scheme of the facial landmarking selection process. Moreover, Figure 8 shows the labels related with every landmark of the shape model. The figure depicts the labels that represents all landmarks of the shape model. Labels $1-17$ correspond to the facial contour; labels $18-27$ correspond to the eyebrows region; labels $28-36$ correspond to the nose region; labels $37-48$ correspond to the eyes, and labels $49-68$ correspond to the mouth region

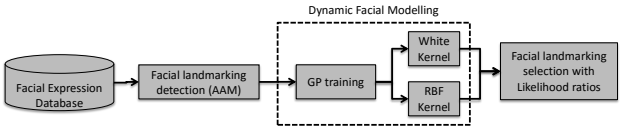


Fig. 7 Dynamic facial landmarking selection process.

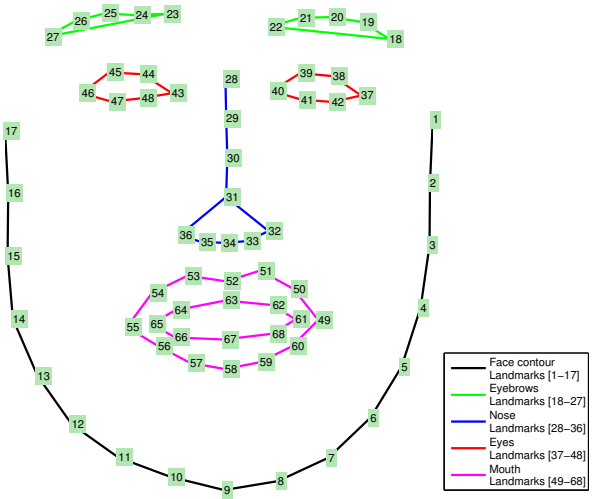


Fig. 8 Landmarks used to define the shape model. The figure shows the set of landmarks related with each face region.

3 Results

3.1 Appearance model estimation error

Table 2, shows the facial landmark detection accuracy. We compute the average of the mean square error between manual landmarks of both databases (CK and FEEDTUM) and the facial landmarks estimated by the AAM model. Also, the standard deviation was computed, as well as the time average of the facial landmarking detection process.

In Table 2, it can be seen that although the accuracy in the facial landmarks detection is greater for images of the training set in CK database, the average error is also small for the test images. This is due to a rigorous procedure in the training of the AAM model, in which we considered facial expressions (emotional process starting from neutral expression to apex[7]) for all prototypical emotions. Moreover, it is noted that although the

[5] we align all emotional time-series using Procrustes analysis.

[6] GPmat is a Matlab implementation of Gaussian processes and other machine learning tools available on `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/software.html`.

[7] Apex is the period during which the emotional expression is maintained at maximal strength.

average error for images with FEED database is a bit higher than in the case of CK database, the accuracy of the estimated model is still higher and fulfills the facial landmark detection task. Also, it is noted that the average time of estimation model (T $[ms]$) is relatively small which would help in on-line applications.

Table 2 Facial features average estimation error. The table shows the robust fit of the facial features, and proves to be applicable in on-line applications.

Set	Database			
	Cohn Kanade		FEEDTUM	
	Error [pix]	T [ms]	Error [pix]	T [ms]
Training	2.0455 ± 0.35	19.1	2.3412 ± 0.41	18.8
Test	2.8577 ± 0.57	19.1	3.1343 ± 0.51	20.3

3.2 Distribution on the relative error

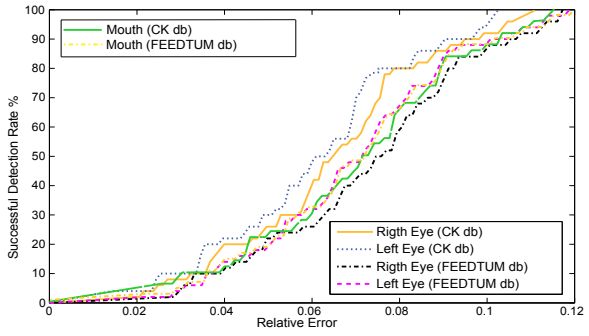
In order to measure the level of matching of the facial landmark to a given expression, we compute the relative error between the manually labeled landmarks and the landmarks points estimated by the AAM model, for eyes and mouth regions respectively. To this end, we compute the euclidean distance for the set of landmarks for each region (distance between manually landmarks and estimated landmarks). Then, we rank all of these distances for all images in the dataset. We follow the criterion of successful detection rate, in which a given estimated contour, corresponds to a plausible region (mouth and eyes respectively) [20]. This criteria establishes that if the relative error, $Rerr = 0.25$, (when the successful detection rate for the euclidean distances reaches 100%), the match of the AAM model to the face is considered to be successful.

Figure 9, shows the distribution function of the relative error against successful detection rate, on which it is observed that for a relative error of 0.1 in the case of the matching of the right eye, 0.098 for the left eye and 0.12 for the mouth region in *CK db* images, the detection rate is 100%, indicating that the accuracy in the matching process of the AAM model is high. The relative error shows the accuracy in which the facial features are estimated in the facial image. Furthermore, it can be seen that the shape model is fitted robustly due to low values of relative errors (less than 0.12), that ensures the correct location of the facial landmarks.

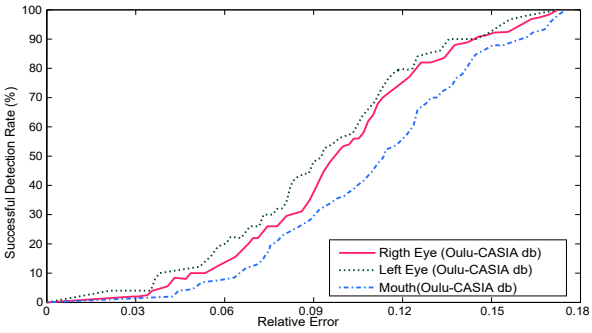
Besides, in Figure 9, it can be seen that relative errors for *FEEDTUM db*, reaches 100% of the detection rate, for relative errors such as 0.118 and 0.119 for the eyes and 0.12 for the mouth region respectively; being these errors much more lower than the established criterion of 0.25.

In order to test our trained model in more complex scenarios, we used the Oulu-CASIA dataset to modeling the facial expression in a given facial image. In addition, Figure 9 shows that when the illumination of the scenario is weak (*Oulu-CASIA db*) the appearance model reaches the 100% of the detection rate, for relative errors between $0.165 - 0.18$ for eyes and mouth regions; which gives us an acceptable error (weak illumination) in comparison with the criterion of 0.25. Moreover, the results show that the trained model can fit accurately a given facial expression even when the illumination scenario is weak.

Furthermore, some other works considered that the criterion of $Rerr < 0.25$ is not suitable for facial features detection in images with lower resolution. Here, relative errors of $Rerr < 0.15$ are considered in order to perform a successful detection [20]. Based on this assumption, we show that the AAM model used in this work is efficient and fulfills this requirement.



(a) Normal indoor ilumination scenarios



(b) Weak ilumination scenarios

Fig. 9 Relative error *vs.* successful detection rate for *CK*, *FEEDTUM* and *Oulu-CASIA* databases.

3.3 Facial landmarking selection

In this section, we present the experimental results using the proposed facial landmarking selection method described above (see section 2.4). Figure 10 shows the

signal noise ratio (SNR) results for both databases, and shows which facial landmarks are more relevant in an emotional sequence. The results show the existence of landmarks that present a high SNR, and can be more relevant in an emotional recognition task. For a quantitative measure of the discriminative landmarks, we perform the GP regression using the best landmarks that represent an emotional sequence. Figure 12 shows the regression process that fits each emotion sequence. This process leads to the best landmarks that represent all emotions. Also, Figure 12 shows these landmarks located in the shape model. The Figure shows the GP regression over an emotional time-series sample in which it can be seen that GPs fits robustly those landmarks that are more relevant in an emotional sequence.

In order to perform a quantitative analysis in the regression process, we compute the effect of the length scale parameter l^2 over the GP. Figure 11 shows that a small length scale means that f varies rapidly along time, and a large length scale, means that f behaves almost as a constant function. Moreover, by using the RBF kernel, we can show that the regression process becomes very powerful when combined with hyperparameter adaptation (see section 2.3).

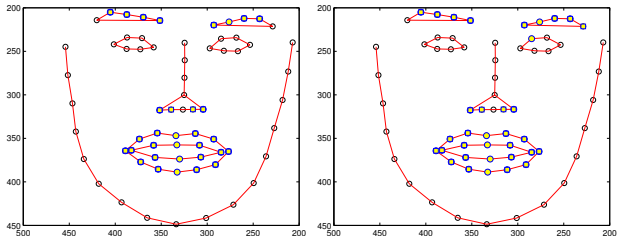
Figure 13, shows the best SNR landmarks located in the shape model for each emotion in all databases. We select the best landmarks as those of who shown higher SNR values in comparison with the average of the SNR value for each emotion. These results show that the facial landmarks more discriminant in emotional sequences are located in mouth, eye and eyebrow facial regions.

To summarize the results derived in the facial landmarking discrimination process, Tables 3 and 4 show the best SNR values for each emotion in the databases used. In both tables it can be seen that the best SNR values correspond to the eyes, eyebrows, and mouth regions (see Figure 8 to find the labels reported in the tables). The results also show that landmarks that exhibit high SNR ratios, can model accurately an emotional process, and would help in affect recognition applications.

3.3.1 Spontaneous Emotions

After studying the emotional sequences on databases in which the emotional process was acted, the RML emotion database was used in order to perform the dynamic facial landmarking selection with likelihood ratios. Here, spontaneous emotional sequences were used to model the facial landmark dynamics. Results are shown in Figure 14. The results show that even for spontaneous emotional sequences, the GP model fits every sequence with high SNR values. Besides, the landmarks ranked in this process correspond with the more relevant landmarks found in the landmark selection process for the other databases (which means that spontaneous emotional process has similar facial expressions with prototypical emotions). Figure 15 shows the best SNR landmark rate for all subjects in the database. The histograms were computed from SNR values from all subjects (analyzing both x-axis and y-axis of the landmarks). We set a threshold by computing the mean of the SNR for all landmarks. Moreover, from this threshold we can establish the SNR level at which a given landmark may be considered relevant in an emotional process.

The main reason for this experiment was to quantify which SNR values (for each landmark), were similar for all subjects in a spontaneous emotional process. The results show that landmarks located in eyebrows, nose-tip and mouth area, are more relevant in a spontaneous emotional sequence.

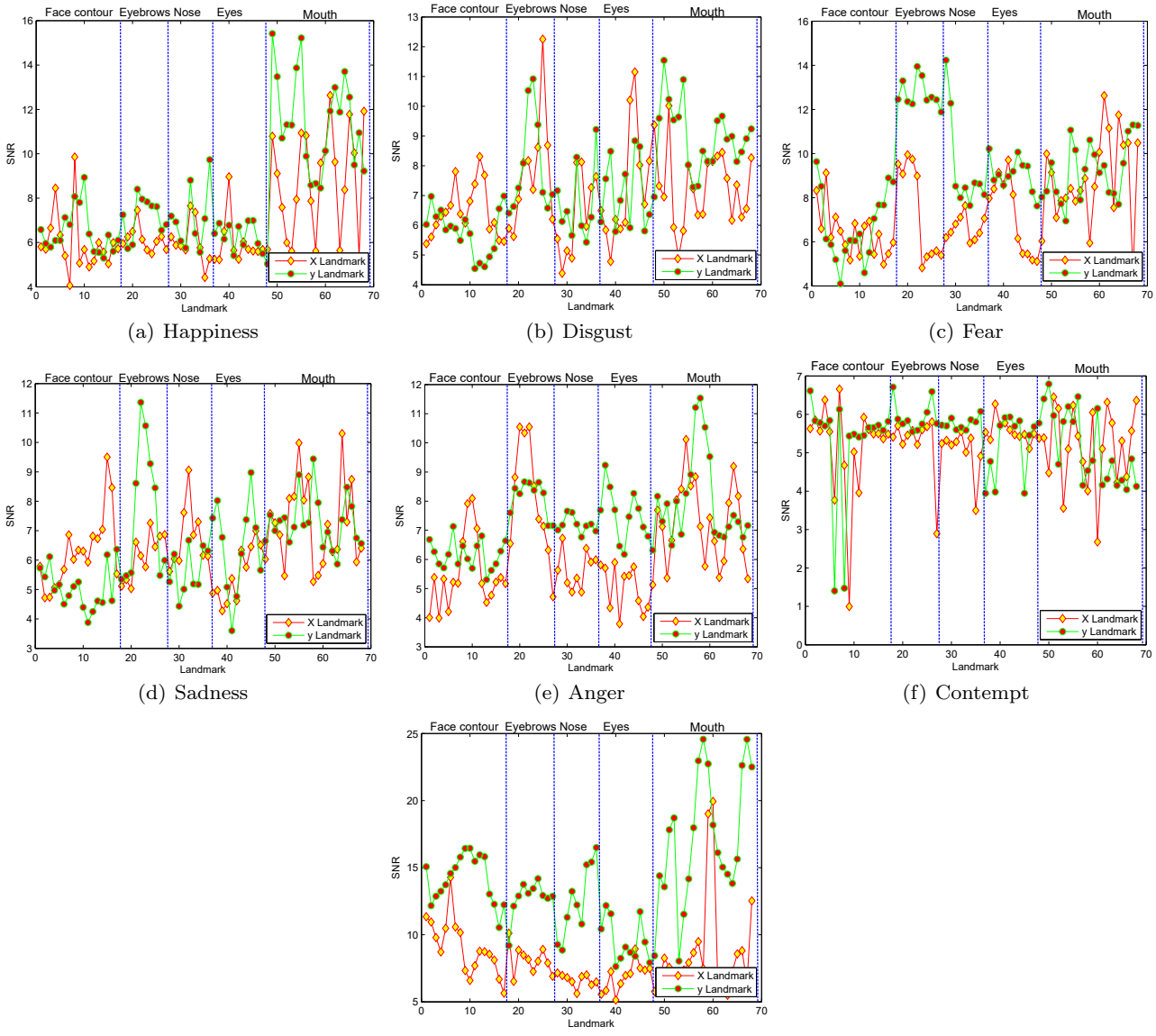


(a) Loglike (b) SNR

Fig. 14 Best landmarks for the spontaneous database. The dynamic facial analysis for the spontaneous emotional sequences proves that an spontaneous emotional behavior shows facial parameters closely related to those facial areas that are more sensitive in the elicitation of the prototypical emotions. Also, Figure 14 shows that landmarks that present more relevance in a spontaneous emotional sequence, correspond to the same landmarks ranked for an emotional acted data-sets (landmarks with high SNR values).

3.4 Emotion recognition using facial landmarks

In order to test our facial landmarking selection method, we evaluate the selected landmarks for each emotional time-series by performing an emotion recognition task for the selected features. To this end, we use Hidden Markov Models (HMM) to perform the emotion recognition. The observations are the time-series for all selected landmarks computed in the ranking process. Tables 5 and 6 show the emotion recognition accuracy for the four databases used in this work (CK, FEEDTUM, Oulu-CASIA and RML). The results show that by using those landmarks selected in the ranking process, the



(a) Happiness (b) Disgust (c) Fear

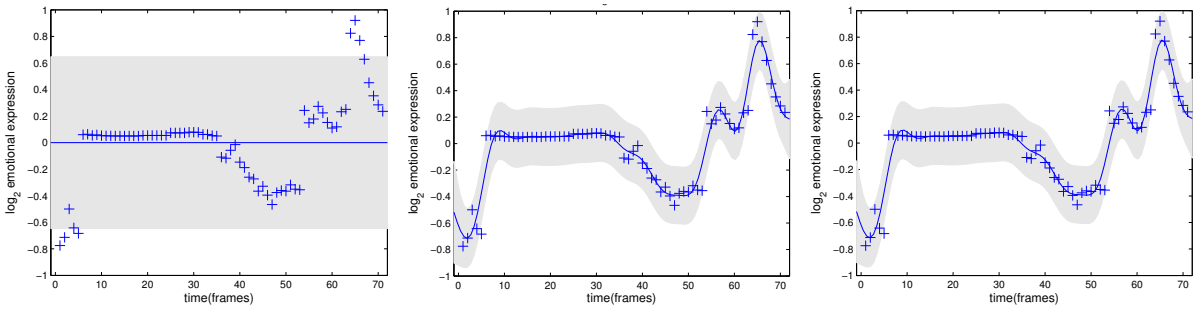
(d) Sadness (e) Anger (f) Contempt

(g) Surprise

Fig. 10 The mean SNR was obtained from every landmark in all emotional sequences (a SNR landmark value for each emotion). Here we compare the white GP variance respect to the RBF variance (radial basis function), to asses the variability of each landmark in an emotional time-series for both x-axis and y-axis respectively. Segments separated by dashed blue lines, represents the set of landmarks for each of the face regions (Face contour, eyebrows, nose, eyes and mouth).

recognition accuracy increases (94.53% for the proposed method with CK database and 93.60 for the FEED-TUM database). Another important result is that the recognition accuracy increases, even when the spontaneous database is used (91.39% for the proposed method with RML database). In addition, the results also show that the recognition rate is accurate even when the illumination scenario is weak (91.75% of accuracy for the Oulu-CASIA database). However, when the model performs the recognition using the entire set of facial features, the recognition rate decreases substantially (see Table 6). The main reason is that when the AAM model fits the facial shape in those scenarios with low illumination, landmarks located in facial contour (i.e. chin landmarks) causes inaccurate recognitions. Besides, the results obtained in this work show that our approach fulfills the results in the state-of-art in emotion recognition tasks in which a given method for facial feature selection is used (see [40], [41], and [33]).

Finally, table 7 shows an experimental comparison between common emotion recognition approaches in the state-of-art. The table shows, that most of the works analyze only 6 prototypical emotions, since contempt emotion has a similar expression that disgust and anger



(a) Emotional expression estimation width $l = 31.6228$ (b) Emotional expression estimation width $l = 8.9443$ (c) Emotional expression estimation width $l = 5.4772$

Fig. 11 Gaussian process regression on an emotional expression in the CK database. The Figure shows, a GP fitted on the emotional expression sample with different settings of the lengthscale hyperparameter l^2 (for a landmark related to a eyebrows region). The blue crosses represent zero-mean facial expression sequence (for a sample landmark) in time (log_2 ratios between facial shape variation) and the shaded area indicates the point-wise mean plus/minus two times the standard deviation (95% confidence region). Figures **(a), (b), (c)** shows different settings of l^2, $l = 31.6228$, $l = 8.9443$ and $l = 5.4772$ respectively.

Table 3 Best mean SNR for each emotion for CK database. The table shows a more descriptive analysis of those landmarks that proved to be more relevant in all emotional sequences of the CK database.

Emotion						
Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
11.212(57)	6.711(18)	11.540(50)	13.946(22)	15.218(55)	10.568(23)	24.559(67)
10.545(20)	6.612(1)	10.920(23)	13.295(19)	13.704(64)	9.980(55)	22.739(59)
10.340(21)	6.461(56)	10.528(22)	12.549(25)	12.987(62)	9.439(58)	22.504(68)
9.527(60)	6.404(49)	10.207(43)	12.442(26)	12.552(65)	9.059(32)	19.009(59)
9.195(65)	6.362(68)	9.668(62)	12.355(20)	11.915(68)	8.898(55)	18.178(60)
8.837(57)	6.271(39)	9.597(49)	12.253(21)	11.778(65)	8.742(66)	17.822(51)
8.663(21)	6.202(54)	9.515(61)	11.747(64)	11.293(53)	8.480(65)	16.448(10)
8.629(22)	6.154(60)	9.371(48)	11.274(68)	10.934(55)	8.455(25)	16.119(61)
8.518(23)	6.076(36)	9.220(36)	11.063(54)	10.791(49)	8.095(53)	15.818(13)
8.436(19)	6.049(59)	8.909(67)	10.615(58)	10.129(60)	8.024(38)	15.642(65)

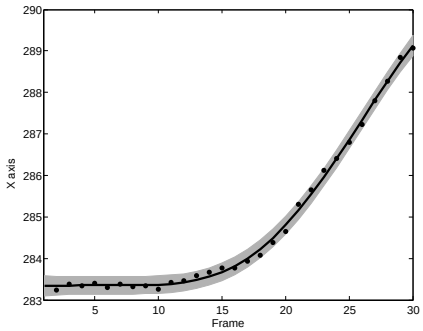
Table 4 Best mean SNR for each emotion for FEED database. The table shows the same extended analysis of the SNR landmarks for each emotion in FEEDTUM database. The results show that the SNR values with this dataset are closely related to the SNR values obtained with the CK database.

Emotion						
Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
11.534(58)	6.790(50)	12.254(25)	14.231(28)	15.415(49)	11.362(22)	24.570(58)
10.548(22)	6.658(7)	11.151(44)	13.535(23)	13.872(54)	10.301(64)	22.967(57)
10.533(59)	6.591(26)	10.896(54)	12.625(61)	13.472(50)	9.501(15)	22.626(66)
10.117(55)	6.454(51)	10.234(51)	12.458(18)	12.634(61)	9.277(24)	19.939(60)
9.231(38)	6.377(4)	10.015(51)	12.423(24)	11.922(61)	8.976(45)	18.709(52)
8.910(56)	6.315(62)	9.638(53)	12.277(29)	11.879(63)	8.825(57)	17.972(56)
8.819(19)	6.233(55)	9.540(52)	11.883(27)	11.314(52)	8.610(21)	16.500(36)
8.645(24)	6.154(52)	9.378(24)	11.302(67)	10.945(67)	8.464(16)	16.433(9)
8.519(56)	6.131(7)	9.240(68)	11.151(62)	10.820(56)	8.162(54)	15.957(12)
8.487(39)	6.053(25)	8.988(64)	11.013(66)	10.705(51)	8.041(56)	15.783(8)

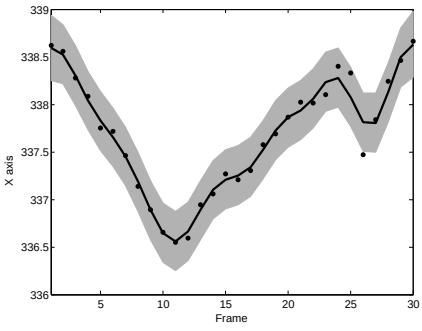
emotions, which makes the recognition method less accurate. Furthermore, the experimental results show that by modeling the temporal behavior of the facial expressions the dynamic features becomes more representative that the static ones (i.e. appearance and geometric features used in works such as [17, 18, 15, 8]).

4 Conclusions and Future Works

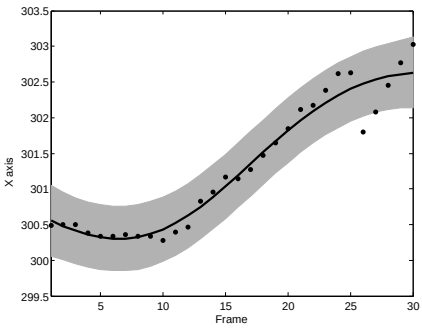
In this paper, we have proposed a method for dynamic facial landmarking selection for emotion recognition by using Gaussian Processes and ranking with log-likelihood ratios. We have shown that the proposed method brings to the state-of-the art, a novel way to analyze which



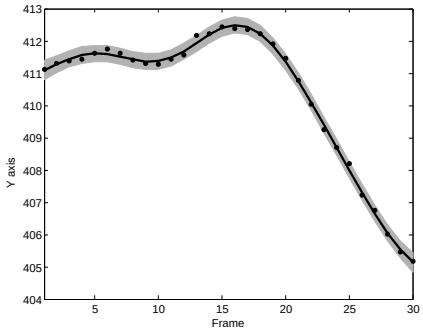
(a) *Landmark* 19, $SNR = 30.2049$ (Anger)



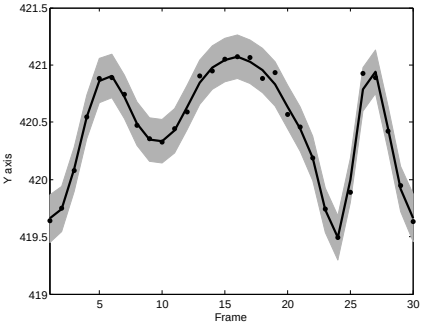
(b) *Landmark* 61, $SNR = 12.9775$ (Contempt)



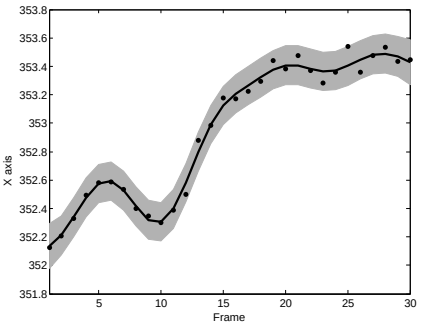
(c) *Landmark* 42, $SNR = 13.0877$ (Disgust)



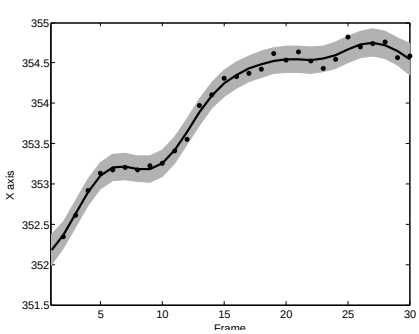
(d) *Landmark* 24, $SNR = 27.0208$ (Fear)



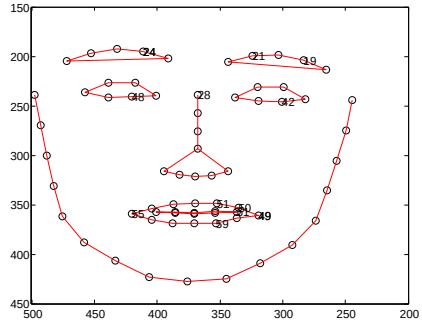
(e) *Landmark* 55, $SNR = 16.6506$ (Happiness)



(f) *Landmark* 51, $SNR = 17.5877$ (Sadness)



(g) *Landmark* 59, $SNR = 20.7552$ (Surprise)



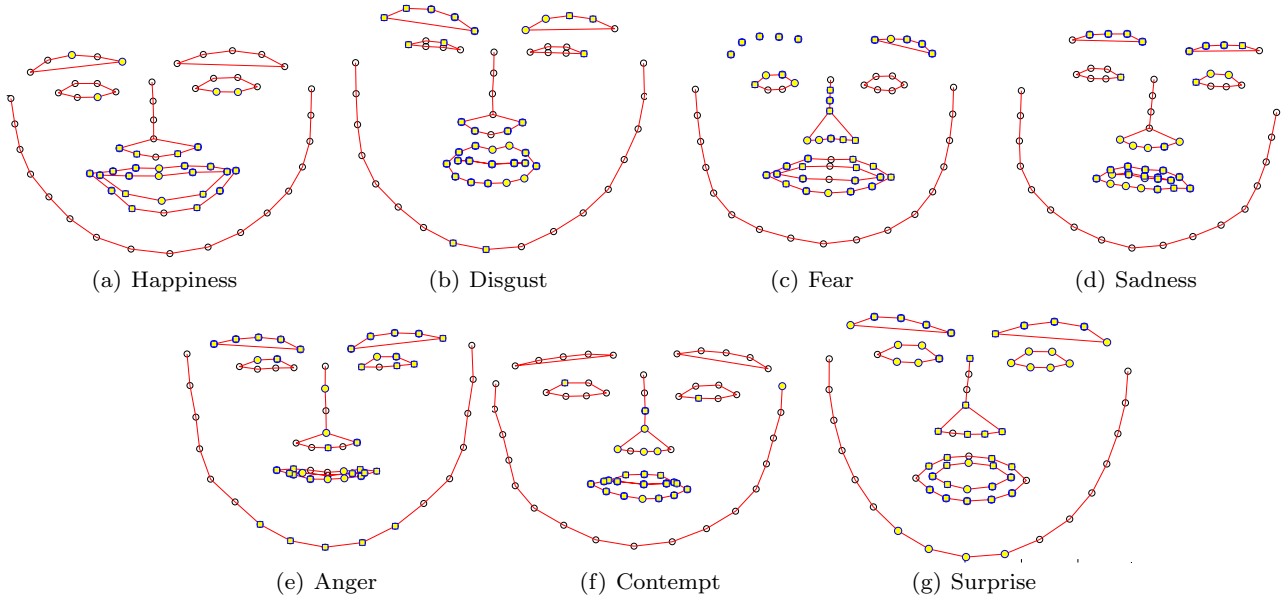
(h) Shape with best SNR landmarks

Fig. 12 GP regression for all emotions. The results show the more relevant landmarks for each emotion (SNR included).

Table 5 Emotion recognition accuracy for the CK, FEEDTUM, Oulu-CASIA and RML databases using the selected landmarks for each emotion. The results show that when the selected facial landmarks are used, the recognition performance increases.

	Database			
Emotion	RML	FEEDTUM	Oulu-CASIA	CK
Anger	90.34 ± 4.56	94.78 ± 2.35	92.68 ± 4.98	95.45 ± 1.29
Contempt	N/A	91.08 ± 3.35	N/A	93.04 ± 3.21
Disgust	89.59 ± 5.64	91.43 ± 3.75	88.73 ± 4.69	92.94 ± 2.77
Fear	91.49 ± 3.87	93.07 ± 2.89	90.46 ± 4.01	94.63 ± 3.02
Happiness	93.48 ± 3.61	96.47 ± 3.22	94.21 ± 3.57	96.73 ± 3.42
Sadness	90.77 ± 4.16	92.95 ± 2.96	91.03 ± 3.78	92.88 ± 2.14
Surprise	92.67 ± 4.03	95.44 ± 1.35	93.42 ± 4.43	96.03 ± 1.46
Average	91.39 ± 4.31	93.60 ± 2.84	91.75 ± 4.24	94.53 ± 2.47

landmarks are more relevant in an emotional sequence. The results show that the facial landmarking detection method is exact and complies with the requests for this type of systems. Through quantitative analysis, the ro-



(a) Happiness (b) Disgust (c) Fear (d) Sadness

(e) Anger (f) Contempt (g) Surprise

Fig. 13 Discriminative landmarks for each emotion (yellow points depict the best ranked landmarks). The results shows that landmarks located in those facial regions that are particular for every emotion (i.e. rise corner of the lips in happiness expression), proves to be more relevant in an emotional sequence.

Table 6 Emotion recognition accuracy for the CK, FEEDTUM, Oulu-CASIA and RML databases using the entire set of facial landmarks. The results show that the entire set of facial landmarks is less representative for emotion recognition processes (low recognition rates).

	Database			
Emotion	RML	FEEDTUM	Oulu-CASIA	CK
Anger	85.46 ± 4.13	91.83 ± 3.89	83.72 ± 4.31	93.01 ± 2.45
Contempt	N/A	87.48 ± 3.64	N/A	91.35 ± 3.41
Disgust	83.72 ± 4.61	89.68 ± 3.92	82.33 ± 3.87	90.87 ± 2.43
Fear	88.52 ± 4.03	90.43 ± 3.45	87.45 ± 4.26	91.79 ± 3.36
Happiness	90.45 ± 3.37	93.21 ± 3.06	91.45 ± 3.86	93.88 ± 4.17
Sadness	87.03 ± 3.32	89.95 ± 3.85	85.46 ± 4.41	90.72 ± 3.06
Surprise	89.06 ± 3.74	90.44 ± 2.35	88.38 ± 3.76	93.42 ± 2.59
Average	89.04 ± 3.87	90.43 ± 3.45	86.46 ± 4.07	92.15 ± 3.07

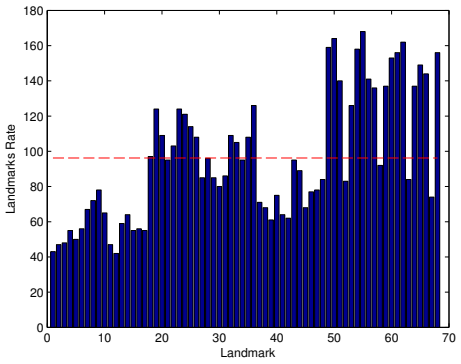
Table 7 Experimental comparison of the proposed method with common state-of-the-art approaches in emotion recognition using CK database.

Approach	**Classes**	**Accuracy** ($mean \pm std$)
Kotsia *et.al.*[17]	6	92.47 ± 4.56
Liu *et.al.* [18]	6	92.33 ± 3.17
Khan *et.al.*[15]	6	78.5 ± 3.05
Chiranjeevi *et.al.* [8]	6	87.67 ± 2.08
Ours	7	$\mathbf{94.53 \pm 2.47}$

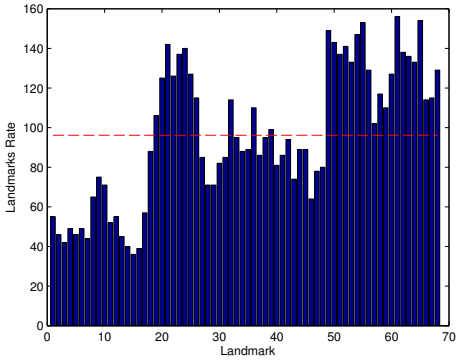
bustness of the AAM model in facial feature detection is evaluated. The results shows that the errors of the AAM model in the matching process remain in nominal values of RMSE (satisfy the criterion of the relative error).

The results shown in the dynamic facial landmarking selection process, prove that supervised learning for regression tasks, offers a robust way to quantify the dynamical facial changes in an emotional sequence. Besides, using GPs to model facial expression time-series, allow us to rank the best SNR landmarks embedded in emotional sequence. Our approach, proves that works developed in the affective computing field, can be improved, since most of these works only use some facial features that belongs to the shape model (i.e. corners of eyes, eyebrows and mouth respectively).

Furthermore, the results show that any emotional sequence exhibits a sets of landmarks that can vary in an given emotional process. Moreover, the proposed method supports works like [30] in which the emotional characterization includes only a few landmarks related to the shape model that are included in the emotion recognition process [30, 3]. In addition, due to high accuracy in the dynamic facial landmarking selection process, the proposed method shows accurate performances for emotion recognition tasks.



(a) x Landmark



(b) y Landmark

Fig. 15 Histograms for best SNR values. In order to perform a quantitative evaluation of the best ranked landmarks, we compute histograms of all SNR values derived from the regression process and so prove that in all emotional sequences some of the facial landmarks bring an important information to the dynamical emotional analysis. Horizontal dashed red line represent the threshold at which a given landmark is considered as relevant. Segments separated by dashed green lines, represent the set of landmarks for each of the face regions.

For future works, we plan to analyze the dynamic changes related to 4D Facial expressions datasets in order to extend our framework for 3D Facial shapes. Moreover, we plan to study, if the facial appearance descriptors (facial landmarks and texture information) present relevant information in an emotional time-series. Finally, we plan to build an emotional ranking process based on multi-output Gaussian process regression framework.

References

1. Alvarez M, Luengo D, Lawrence N (2013) Linear latent force models using gaussian processes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(11):2693–2705, DOI 10.1109/TPAMI.2013.86
2. Bishop CM (2007) Pattern Recognition And Machine Learning (Information Science And Statistics). Springer, URL `http://www.openisbn.com/isbn/9780387310732/`
3. Bousmalis K, Mehu M, Pantic M (2013) Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. Image Vision Comput 31(2):203–221, DOI 10.1016/j.imavis.2012.07.003, URL `http://dx.doi.org/10.1016/j.imavis.2012.07.003`
4. Carl Edwards Rasmussen CW (2006) Gaussian Processes for Machine Learning. The MIT Press
5. Chakraborty A, Konar A, Chakraborty U, Chatterjee A (2009) Emotion recognition from facial expressions and its control using fuzzy logic. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 39(4):726 –743
6. Cheon Y, Kim D (2008) A natural facial expression recognition using differential aam and knns. In: Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia, IEEE Computer Society, Washington, DC, USA, ISM '08, pp 220–227
7. Cheon Y, Kim D (2009) Natural facial expression recognition using differential-aam and manifold learning. Pattern Recogn 42:1340–1350
8. Chiranjeevi P, Gopalakrishnan V, Moogi P (2015) Neutral face classification using personalized appearance models for fast and robust emotion detection. IEEE Transactions on Image Processing 24(9):2701–2711
9. Ekman P (2007) Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life, 2nd edn. Owl Books, 175 Fifth Avenue, New York
10. Ekman P, Friesen W (1978) Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto
11. Ekman P, Rosenberg E (2005) What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford Univ. Press
12. Gunes H, Pantic M (2010) Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Proceedings of the 10th international conference on Intelligent virtual agents, Springer-Verlag, Berlin, Heidelberg, IVA'10, pp 371–377
13. Jack RE, Garrod OG, Schyns PG (2014) Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Current Biology 24(2):187 – 192, DOI http://dx.doi.org/10.1016/j.cub.2013.11.064,

URL http://www.sciencedirect.com/science/article/pii/S0960982213015194
14. Kalaitzis AA, Lawrence ND (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. BMC bioinformatics 12(1)
15. Khan RA, Meyer A, Konik H, Bouakaz S (2011) Facial expression recognition using entropy and brightness features. In: 2011 11th International Conference on Intelligent Systems Design and Applications, pp 737–742
16. Kirk PDW, Stumpf MPH (2009) Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. Bioinformatics 25(10):1300–1306
17. Kotsia I, Buciu I, Pitas I (2008) An analysis of facial expression recognition under partial facial image occlusion. Image Vision Comput 26(7):1052–1067
18. Liu Q, Metaxas DN, Yang P (2010) Exploring facial expressions with compositional features. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 00:2638–2644
19. Lucey P, Cohn J, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) 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, pp 94–101
20. M H, Ido S (2009) Eye detection using intensity and appearance information. Springer-Verlag pp 801–809
21. MacKay DJC (2002) Information Theory, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA
22. Matthews I, Baker S (2004) Active appearance models revisited. Int J Comput Vision 60:135–164
23. Murphy KP (2012) Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning Series). The MIT Press, URL http://www.openisbn.com/isbn/9780262018029/
24. Nicolaou M, Gunes H, Pantic M (2011) Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. Affective Computing, IEEE Transactions on 2(2):92–105, DOI 10.1109/T-AFFC.2011.9
25. Nicolaou M, Gunes H, Pantic M (2011) Output-associative rvm regression for dimensional and continuous emotion prediction. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp 16 –23
26. Pantic M, Patras I (2005) Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In: Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics, pp 3358–3363
27. Pantic M, Patras I (2006) Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans Systems, Man, and Cybernetics, Part B 36:433–449
28. Pun T, Pantic M, Soleymani M (2012) Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing 3(2):211–223, DOI http://doi.ieeecomputersociety.org/10.1109/T-AFFC.2011.37
29. Rasmussen CE, Williams CKI (2005) Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press
30. Rudovic O, Pantic M, Patras I (2013) Coupled gaussian processes for pose-invariant facial expression recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(6):1357–1369
31. Shuai-Shi L, Yan-Tao T, Dong L (2009) New research advances of facial expression recognition. In: Machine Learning and Cybernetics, 2009 International Conference on, vol 2
32. 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, Springer-Verlag, Berlin, Heidelberg, ECCV '08, pp 58–71
33. Taheri S, Qiu Q, Chellappa R (2014) Structure-preserving sparse decomposition for facial expression analysis. IEEE Transactions on Image Processing 23(8):3590–3603, DOI 10.1109/TIP.2014.2331141
34. Valstar M, Pantic M (2012) Fully automatic recognition of the temporal phases of facial actions. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 42(1):28–43
35. Wallhoff F (2006) Database with Facial Expressions and Emotions from Technical University of Munich (FEEDTUM). URL http://www.mmk.ei.tum.de/\~waf/fgnet/feedtum.html
36. Wang Y, Guan L (2008) Recognizing human emotional state from audiovisual signals. IEEE Transactions on Multimedia 10(4):659–668
37. Wu CH, Lin JC, Wei WL (2013) Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course. Multimedia, IEEE Transactions on 15(8):1880–1895, DOI 10.1109/TMM.2013.2269314
38. Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: Audio, vi-

sual and spontaneous expressions. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(1):39 –58

39. Zhao G, Huang X, Taini M, Li SZ, Pietikäinen M (2011) Facial expression recognition from near-infrared videos. Image and Vision Computing 29(9):607 – 619, URL http://www.sciencedirect.com/science/article/pii/S0262885611000515
40. Zhao K, Chu WS, la Torre FD, Cohn JF, Zhang H (2016) Joint patch and multi-label learning for facial action unit and holistic expression recognition. IEEE Transactions on Image Processing 25(8):3931–3946, DOI 10.1109/TIP.2016.2570550
41. Zhong L, Liu Q, Yang P, Huang J, Metaxas DN (2015) Learning multiscale active facial patches for expression analysis. IEEE Transactions on Cybernetics 45(8):1499–1510, DOI 10.1109/TCYB.2014.2354351