International Journal of Advanced Computer Science, Vol. 3, No. 5, Pp. 202-216, May, 2013.

Continuous Facial Expression Representation for Multimodal Emotion Detection

Catherine Soladié, Hanan Salam, Nicolas Stoiber & Renaud Seguier

Manuscript
Received:
12,Mar.,2013
Revised:
25,Mar.,2013
Accepted:
1, Apr.,2013
Published:
15,Apr.,2013

Keywords
Facial Expression Representation, Transfer learning, Fusion Techniques, Fuzzy Inference System, Radial Basis Function (RBF), Context in Emotion Recognition

***Abstract*— This paper presents a multimodal system for dimensional emotion detection, that extracts and merges visual, acoustic and context relevant features. The paper studies the two main components of such systems: the extraction of relevant features and the multimodal fusion technique. Additionally, we propose a method for the automatic extraction of a new emotional facial expression feature to be used as an input of the fusion system. The feature is an invariant representation of facial expressions, which enables person-independent high-level expression recognition. It relies on 8 key emotional expressions, which are synthesized from plausible distortions applied on the neutral face of a subject. The expressions in the video sequences are defined by their relative position to these 8 expressions. High-level expression recognition is then performed in this space with a basic intensity-area detector. In this paper, the impact of fusion techniques is investigated by comparing two different fusion techniques: a fuzzy inference system and a radial basis function (RBF) system. The experiments show that the choice of the fusion technique has little impact on the results, thus indicating that the feature extraction is the key component of a multimodal emotion detection system. The experiments have been performed on the AVEC 2012 database.**

1. Introduction

Our aim is to develop a method to automatically detect when changes in emotion occur. Among the various signals that can be used to detect emotions, visual and acoustic features play an important part. The context in which the video sequence is recorded also contains key information.

Two main parts are necessary in such an automatic system: the extraction of multimodal features from one hand, the fusion system on the other hand.

The way emotions are described is also a key issue of such systems. Even if there is still no agreement about the way an emotion have to be described, continuous representations seem to gain the upper hand. To take into account the complexity of the emotional state, the way emotions are described has recently evolved from a prototypal modelling approach to a dimensional approach [11]. In 2011, the first Audio/Visual Emotion Challenge and Workshop (AVEC2011 [25]) proposed to compare multimedia processing and machine learning methods for automatic audiovisual emotion analysis. The database has a large collection of audio-video sequences, displaying conversations between an emotional agent and an unknown subject. In this challenge, the emotion is labelled in terms of positive/negative valence, and high and low arousal, expectancy and power. In 2012, we participated in the second edition of this challenge (AVEC2012 [24]), and arrived second [30]. This time, the emotion is labelled continuously in the 4 dimensions: valence, arousal, power and expectancy [11].

In this paper, we present a multimodal system for dimensional emotion detection, that extracts and merges visual, acoustic and context relevant features. We precise the extraction of the various multimodal features, that we used in the challenge AVEC 2012 [30] and we more precisely focus on the recognition system of high level facial expressions.

In addition, we investigated the impact of the fusion system on the overall process. To answer this question, we propose to compare the fusion system we used in the challenge [30] to the fusion system the winners of the challenge used [20] (the other competitors were further behind). We therefore apply those two methods of fusion to merge the same high-level features: a fuzzy inference system and a radial basis function system.

Before presenting the systems, we will briefly present a state-of-the-art on multimodal emotion recognition systems and on facial expression recognition including the components of this system (facial features extraction).

This work was supported by Immemo (French ANR project) and Replica (ANR Techsan).

Catherine Soladié, Hanan Salam, Renaud Seguier are with Supelec/IETR, Avenue de la Boulaie, 35576 Cesson-Sévigné, France.
(✉ catherine.soladie@supelec.fr, hanan.salam@supelec.fr, renaud.seguier@supelec.fr).
Nicolas Stoiber is with Dynamixyz, 80 avenue des Buttes de Coesmes, 35700 Rennes, France (nicolas.stoiber@dynamixyz.com).

A. Multimodal Emotion Recognition

Multimodal emotion recognition has developed in recent years [39].

1) Fusion system

The fusion of the modalities can be done at different stages. In early data fusion, that is the fusion of the features before the recognition process, the features of the various modalities are either directly concatenated [2] or the correlation between the features of the different mode is taken into account. This is done for instance by HMM methods [31], neural network methods [12] [14] or Bayesian network methods [26]. In late data fusion, that is the fusion of the recognition results of the various modalities, the recognition results are fused with for example empirical weights [13] or rules [22]. In this paper, we proposed two fusion methods using early data fusion: a fuzzy inference system and a radial basis function prediction system.

2) Modalities

Most systems merge data from the two modalities: acoustic (prosody) and visual (facial expressions). Context or lexical indicators [23] are rarely taken into account. Context information can be specific to the database, for example the fact that the more subjects are engaged into a conversation, the more intense their emotions might get [21].

3) Features level

One of the key points in emotion recognition systems is the level of the features used for the prediction. Some methods used low-level features [20] [23]. These systems often suffer from the size of the data to process. On the contrary, systems using high-level features [21] have to deal with fewer features and often have the advantage to give a meaningful relation between a feature and the emotion.

4) Emotional labels

The majority of the systems focus on the classification of discrete emotions [2] [13] [26]. Some systems evolved towards a dimensional representation of emotions (activation, valence, evaluation) but the output value remains discrete [14]. With AVEC2012 challenge and workshop, the emotional labels are in four continuous dimensions (valence, arousal, expectation and power) [20] [21] [23].

B. Facial Expression Recognition

Many expressions recognition systems have been proposed in the last decade [10] [33]. Many of them focus on the classification of facial expressions into 6 categories corresponding to the 6 emotions universally associated with distinct facial expressions [8]. Few detect other expressions such as pain [16] [1] or deal with the intensity of the expression [9].

The choice of representation is known to influence the recognition performance. Most systems directly use the geometric and/or appearance features (shape and/or texture features). Other systems use an intermediary representation before facial expression recognition. For instance, [7] extract Action Units defined in the FACS. These have the advantage of removing the person-specific characteristics such as the identity or the morphology of the person.

In the last few years, some continuous representations of facial expression have been proposed, using manifold learning [3] [28] [4] [32]. The aim is then to represent the whole facial expression space.

As for the multimodal analysis of emotions, challenges were organized to compare the methods on identical databases. The last one, Facial Expression Recognition and Analysis challenge (FERA 2011 [35]) consisted of the recognition of discrete emotion won by [37] [36] and detection of AUs that our team won [27] with the ISIR laboratory. Even if the results were encouraging, the recognition rates remained low: the person-independent discrete emotion recognition did not exceed 75.2% (although the person-specific performance was 100%) and the AU detection only reached 62 %.

C. Facial Features Extraction

Whether using an intermediary representation or not, the choice of the facial features to extract is a crucial step for a robust recognition system. Two types of features are usually extracted: appearance and geometric features. The formers are ones that are concerned with texture representation such as the use of LGBP [19]. The latters are those that extract information regarding the shape of the facial components (mouth, eyebrows and eyes). Active Appearance Models [5] are statistical tools that can represent the face by a set of ‘appearance parameters’. These parameters encode both the shape and the texture of the subject, which shows their interest.

Extracting features using AAMs can be done in three ways: either through the use of a global model that encodes the whole face, or through local models where areas of interest are encoded separately, or through the combination of local and global models (usually called the hierarchical approach). These approaches benefit from both the generality of a global AAM and the precision of local models in their corresponding areas. This can help to improve the performance of an expression recognition system.

Among the various hierarchical methods in the literature, [40] proposed the component based AAM. In this approach, sub-models iteratively update component points independently and are then united to a global AAM. [15] proposed an approach which automatically identifies independent distinct entities of the face called "cliques". Each of these entities is then modelled using AAM. Finally, a global AAM is created by regrouping the close cliques two-by-two. Even though this method is efficient for dealing with the non-linearities of AAM, it is unable to be

applied to faces with varying poses. [38] modelled the face by a two-level hierarchical person-specific model. The first level accounts for the low level component facial actions (mouth and eyes). The second one combines the facial sub-models to model the final facial expression using intrinsic functionalities (expression variabilities) of sub-components. [34] combined local and global models based on Markov Random Fields, which models the spatial relationships between points. Line segments between points are used as the basic descriptors of the face shape. The local models efficiently select the best candidate points, while the global model regularizes the result to ensure a plausible final shape. [36] employed a multi-resolution grammatical face model. Their model is composed of two layers. The first refines the global AAM model with a set of local models for the different components of the face (eyes, eyebrows, nose and mouth). The second refines the component local models by adding sketch curves to model possible skin marks and wrinkles. In this paper, we propose another way of taking benefits of both local and global models.

D. Overview of our Approach

In this paper, we propose a fully automatic system that extracts relevant features for spontaneous affective audio-video sequences and computes the potentially felt emotional state of the unknown subject.

The article focuses on facial expressions detection. The main contribution of the article is that our system passes from an appearance space to an emotion space through the use of an intermediate expression space, which takes into account the intensity of the expressions, and is invariant across subjects. The particularity of the method is that we did not focus on the appearance space, which carries morphological information of the subject, but on the organization of the expressions with respect to each other. This organization is invariant across the subjects. A facial expression can then be detected with a basic intensity-area detector in this expression space. Another main contribution is that that our system adapts to the subject. Its originality is that the neutral face is automatically computed by the mean value of the appearance parameters of the video sequence and that known plausible distortions are applied on this neutral face to create a person-specific appearance space.

The computation of the appearance space of our system is done through AAMs. We contribute at this level by proposing a hierarchical model called the Multi-Model AAM. This model extrinsically combines a local model of the mouth and a global one of the face thus gaining the merits of global and local alignment.

Another important contribution of the article is the comparison of two fusion techniques on the same relevant high-level features: the fuzzy inference system used in [30] and a radial basis function (RBF) system inspired from [20]. The aim is to evaluate the impact of the fusion method on the global process.

The remainder of this paper is organized as follows. In section 2, we describe the global process for multimodal emotion recognition. In section 3, we define the relevant features extraction. Section 0 focuses on the facial expression extraction. Section 5 presents the feature selection and two fusion methods. Section 6 shows and discusses the results. Section 7 concludes the paper.

2. Global Process

This section presents the global system for emotion detection (see Fig. 1). Two fusion systems are compared: a fuzzy inference system (see subsection 2.B) and a radial basis function system (see subsection 2.C). To be compared, both systems take in input the same relevant features that result from emotional states (see subsection 2.A). The output is a continuous prediction of 4 emotional dimensions: valence, arousal, power and expectancy.

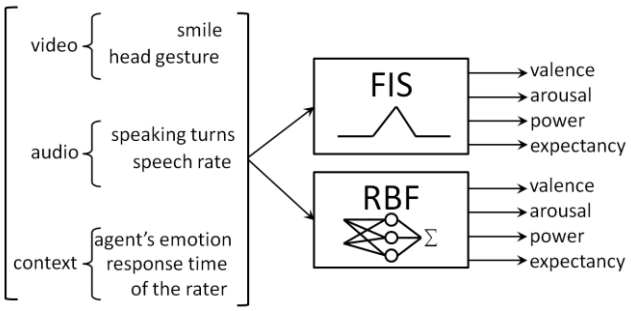


Fig. 1 Overall view of the proposed methods: two fusion systems (a fuzzy inference system and a radial basis function system) transform the relevant features from video, audio and context into 4 emotional dimensions.

A. Feature Extraction Layout

The relevant features extraction is described in subsection 3. A cross-correlation between each feature and each ground truth emotional dimension is performed to select which feature is relevant for each emotional dimension (see section 5.A).

B. A multimodal Fuzzy Inference System Layout

The fuzzy inference system takes in input the relevant feature extracted. Fuzzy rules are defined from the analysis of the data of the training and development databases (see section 5.B). The output information is defuzzified to output the 4 emotional dimensions.

C. A Radial Basis Function Layout

The radial basis function system is inspired from [20]. It takes in input the relevant feature extracted. A set of representative samples of input relevant features is computed via k-means clustering algorithm (see section 0), and are used as the center of radial basis function, for emotion prediction. This process is performed for each of the 4 emotional dimensions.

3. Relevant Features Extraction

The choice of the features to extract has been done according to the analysis of the ground truth emotional labels of the training and development databases (annotations of raters using FEELTRACE [6]). The chosen features were those that explain the global trend of the emotion dimensions and not the small subtle variations of the emotions. They can be classified in 3 categories:

- video features, that include facial expressions (especially laughter) and body language (especially head gesture);
- audio features, that include speaking turns and speech rate;
- context features, that include the emotional agent of the conversation (agent's name or emotional words that are said during the conversation), the response time of the rater and the conversation time.

The features extraction is made from 3 different data sources: videos, speech transcripts and ground truth labels (see Fig. 2). The extraction is described thereafter for each source.

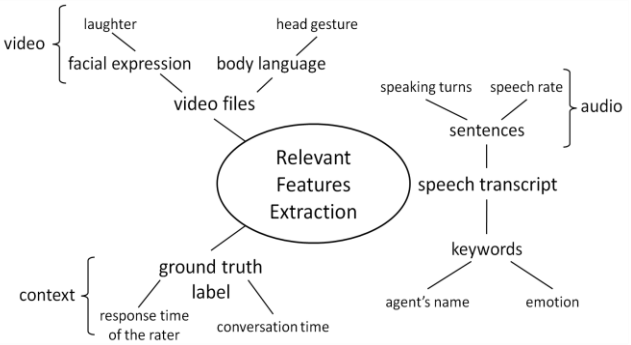


Fig. 2 Sources of the relevant features: video files, speech transcripts and ground truth labels.

A. From Video Files

The extracted features from video files are the facial expressions and the body language.

As the main contribution of this article is about facial expressions, a precise description of the extraction of facial expressions is presented in a dedicated section (section 0).

Concerning the body language, we computed the global movement of the head pose in the scene. The video data are analysed using a person-independent AAM [5] built on the training and development databases. In the test phase, the pose parameters of the face are computed from the AAM model. The body movement is computed from the standard deviation of the head pose in a video sequence with a sliding temporal window of 40 seconds. The more the subject moves and makes wide movements, the higher this quantity is.

B. From Speech Transcripts

The analysis of the sentences gives the length of the sentences pronounced by the subject. In our system, we use binary information. For each speaking turn in a conversation, if the number of words pronounced by the subject is high (above 35 words, empirical data learnt on training and development databases), the sentence is long; otherwise, the sentence is short.

The analysis of the sentences also gives the speech rate. The speech rate is computed from the transcripts by the rate: number of words by time unit.

The conversations are performed between a subject and an emotional agent, who is set in one of the four quadrants of arousal-valence space (Spike is aggressive; Poppy is cheerful; Obadiah is gloomy; and Prudence is pragmatic). We perform a statistical analysis on the sequences for each emotional agent. Fig. 3 shows that the emotional state displayed by the subject matches the one displayed by the agent. For example, if the agent is Poppy, then the subject speaking to Poppy has a tendency to display behaviours of high valence and high arousal. Fig. 4 illustrates this observation with the example of the 4 conversations of subject 2. To find automatically who the emotional agent of the sequences is, we extract names from keywords. They provide some contextual information on which emotional agent the subject is speaking to. Indeed, at the beginning of each conversation, subjects select the agent they want to interact with by telling its name; at the end of their conversation, they select the next emotional agent for their next interaction. When the name of the agent is not pronounced, the subjects usually use emotional terms, such as 'angry' or 'annoy' for 'Spike', 'sad' for 'Obadiah' and 'fun' for 'Poppy'. Thus, the keyword 'angry' or 'Spike' appears in conversation with Spike. It is then possible, with the transcripts of a conversation, to automatically find the emotional agent of the sequence, and to deduce a statistical mean value of the subject's valence and arousal for the sequence.

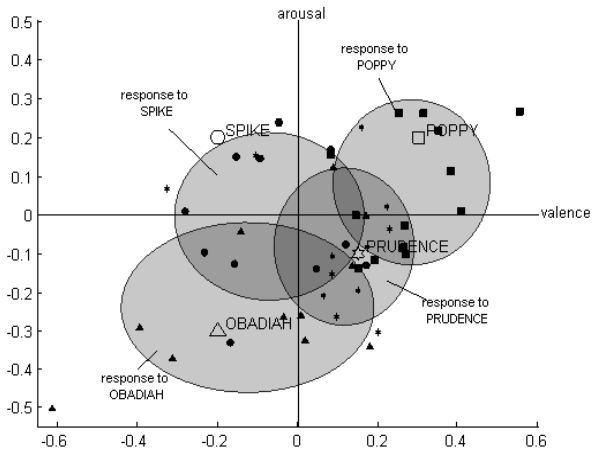


Fig. 3 Displayed emotional response of subjects when interacting with agent. The emotion of the agent is showed by an empty marker (Spike is aggressive; Poppy is cheerful; Obadiah is gloomy; and Prudence is pragmatic). The mean ground truth labels of each subject responding to these agents are showed by a filled marker and their distributions are in the ellipses (mean value and standard deviation).

C. From Ground Truth Labels

The analysis of the ground truth labels highlights a delay in the start of annotations. We computed the mean value and standard deviation of the training and development

ground truth labels for each emotional dimension. We also extracted the values of the ground truth labels in the beginning of the sequences. First we noticed that the labels in the beginning of the sequences are identical for all the sequences; and secondly that they are very different from the values inside the conversation (see Fig. 5) for arousal and power.

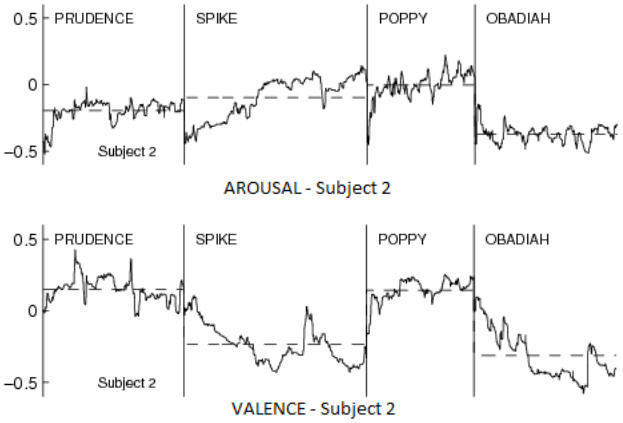


Fig. 4 Ground truth emotional labels (solid line) compared to the mean value of the emotional labels among all the subjects speaking to the emotional character (dotted line). Example of the 4 conversations of the subject 2. Arousal on the first graph, valence on the second graph.

This may be due to the initialization of the tool used to rate and to the response time of the rater, so that the first seconds of the ground truth labels may not be representative of the subject's emotion. Nevertheless, for the challenge, we modelled this behaviour with a feature as a decreasing linear function on the first 20 seconds of the conversation.

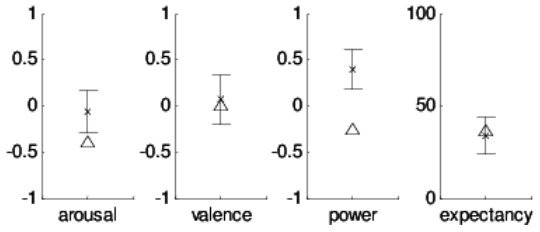


Fig. 5 Impact of the response time of the raters on the ground truth labels: the triangle shows the value of the label at the beginning of the sequence and the cross shows the mean and standard deviation of the labels, for each dimension.

Finally, the analysis of the labels also highlights that the expectancy varies quite similarly across conversations over the time. In the beginning of the conversation (first minute), the expectancy is low. Then, the expectancy is higher. We modelled this behaviour with a square-wave signal (high value the first minute, low value otherwise).

4. Facial Expression Extraction

This section presents the main contribution of the article. After a brief description of the overall process (section 4.A), each main step of the facial expression extraction is described (sections 4.B, 4.C, 4.D and 4.E).

A. Overview of the Process

The global overview of the process is presented in Fig. 7. The main idea is to take into account the morphology of the subject. The process is composed of 4 steps. The first step concerns the detection of the features of the face by a person-independent AAM (section 4.B). The second step computes a person-specific appearance space (section 4.C). The third step transforms this appearance space into a person-independent expression space (section 4.D). The last step performs the expression recognition (section 4.E).

B. Multi-Model AAM

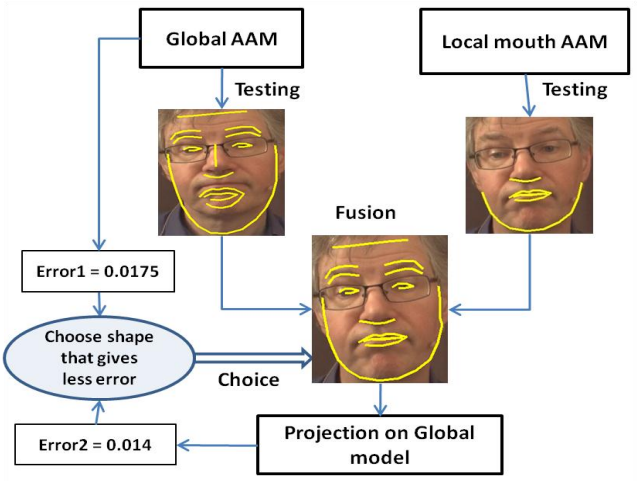


Fig. 6 Example of person-independent Multi-Model AAM (MM-AAM).

The shape of each face image of each video sequence is extracted using Active Appearance Models (AAM) [5]. For the training and development databases [24], we use a Global Face AAM (GF-AAM) which is trained on some images of these two. Regarding the test database, the presence of hair on the face for some persons misleads the GF-AAM in finding a precise localization of the mouth. On the other hand, for other persons in the database (person with beard), a local model fails while the global model does not. So, we propose the Multi-Model AAM (MM-AAM) (cf. Fig. 6) for this database. This MM-AAM combines the results of a Local Mouth AAM (LM-AAM) (trained on the same images as the GF-AAM) and the GF-AAM. The best shape (between the GF-AAM and MM-AAM) is obtained by computing projection errors on the same global AAM. This permits to take advantage of the precise localization of the mouth by LM-AAM when there is hair covering the face and the ability of the GF-AAM to generalize to new faces by using the correlations between the different parts of the face for the other cases.

The algorithm is the following:

1. Train both models: GF-AAM and LM-AAM;
2. Apply both models on the testing videos: Get the global face shape S_{GF} and the local mouth shape S_{LM};
3. Substitute mouth shape from the LM-AAM in the shape from the GF-AAM: get the Multi-Model shape S_{MM};

4. Project S_{MM} on the GF-AAM to obtain the corresponding appearance parameters and the projection error:
   a. Align the S_{MM} to the mean shape $\overline{s}$ of GF-AAM;

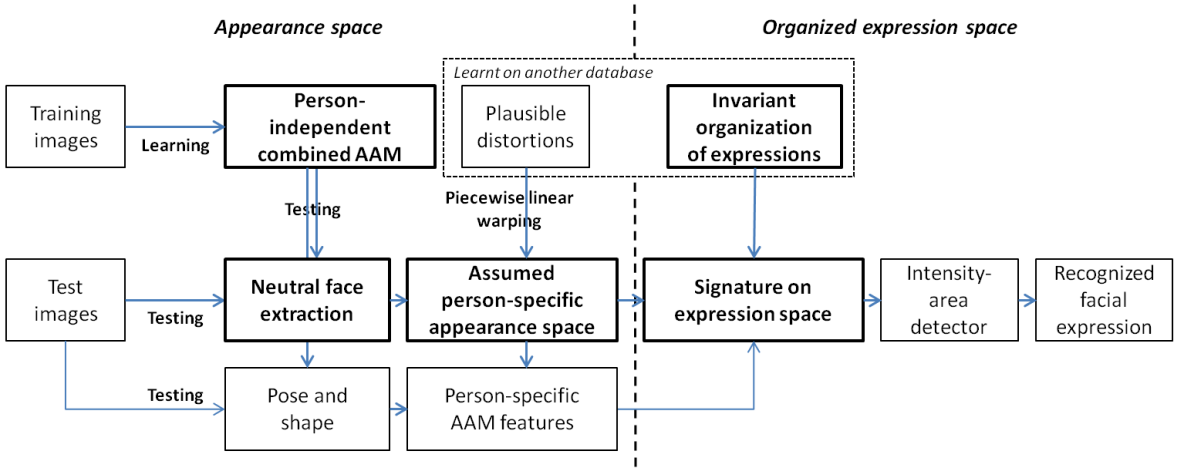


Fig. 7 Overall view of the facial expression extraction. The neutral face of each subject and the shape of all the images are extracted using a person-independent combined AAM. An assumed person-specific appearance space is created by applying plausible distortions on the neutral face of the given subject. The person-specific appearance space is transformed into the expression space using an invariant organization of expressions.

b. Find the shape parameters b_s of S_{MM} using $s = \bar{s} + V_s b_s$. V_s is the matrix build with the shape eigenvectors of the GF-AAM;
c. Warp the texture under S_{MM} into mean shape $\bar{g}$;
d. Find the texture parameters b_g using $g = \bar{g} + V_g b_g$. V_g is the matrix build with the texture eigenvectors of the GF-AAM;
e. Concatenate b_s and b_g: $\begin{pmatrix} W_s b_s \\ b_g \end{pmatrix}$. W_s is the weighting between pixel distances and intensities.
f. The projected appearance parameters are then: $c = V_c b$

5. Choose the shape (S_{MM} or S_{GF}) that gives the lowest projection error defined as the difference between the model synthesized image using the appearance parameters and the texture of the real image defined by the shape.

Confidence extraction: After extracting the shapes of all the frames of the videos, each frame is given a binary confidence index. The latter is computed based on the analysis of projection errors of samples of the sequence in question. As a matter of fact, a threshold error is set for every sequence. If the error of one frame is less than or equal to this threshold error, then the frame is considered to have a good alignment and thus is given a confidence index of 1, else it is assigned a confidence index of 0.

C. Assumed Person-Specific Appearance Space

Our system adapts to the morphology of each subject by creating a person-specific appearance space from the neutral face and plausible expressions.

1) Neutral Face Extraction:

The neutral face of a subject is automatically extracted from the video sequences of this subject. The extraction is made by computing the mean value of the appearance parameters of the person-independent AAM when the subject is not speaking. The neutral face is the image that has the closest appearance parameters from this mean value. Indeed, with a person-independent AAM, the appearance parameters carry both information of the expression and the morphology. The main distortions are dues to the various morphologies of the subjects rather than to the variations due to the expressions (see figure Fig. 8). Fig. 9 shows some examples of neutral faces.

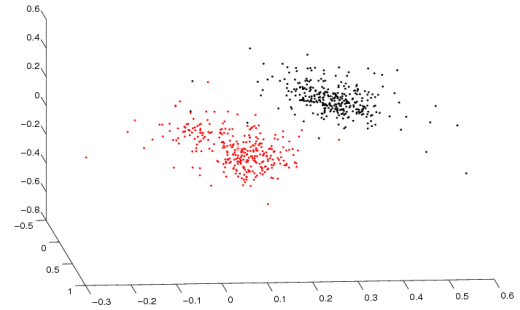


Fig. 8 Three first components of appearance parameters of two subjects (in black and red dots) with a person-independent AAM. The main distortions are dues to the various morphologies of the subjects rather than to the variations due to the expressions.

Fig. 9 Examples of neutral faces extracted from the appearance parameters of the person-independent combined AAM. The neutral face is the closest image to the mean value of the appearance parameters when the subject is not speaking.

2) Assumed shape model:

To create a person-specific appearance space, we use 8 plausible expressions of the subject and we compute a person specific shape model by applying PCA on these 8 plausible expressions plus the neutral face. Plausible expressions are computed from known distortions applied on the neutral face of the subject (see Fig. 10). Each point of each expression is transferred on the neutral face of the subject by piece-wise affine warping. We use 8 known distortions, learnt on another database[1]. Each distortion corresponds to a specific emotional facial expression.

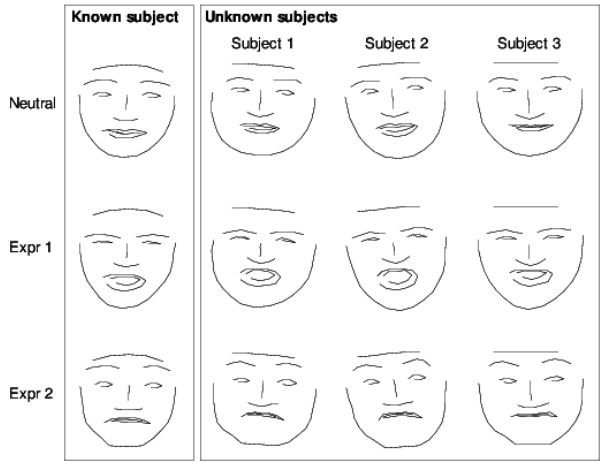


Fig. 10 The plausible distortions are learnt from a known subject and applied on the neutral face of the unknown subjects. (Neutral) Neutral face of the known subject and of 3 unknown subjects. (Expr x) Real expression of the known subject and plausible expression of the unknown subject.

D. Organized Expression Space of an Unknown Subject

To perform expression recognition, the person-specific appearance space is transformed into a person-independent expression space. This transformation is performed by using an invariant representation of facial expressions. Instead of describing one expression by its appearance features (which means taking into account the morphology of the subjects), we describe one expression by its relative position to others. We previously showed [29] that the organization of 8 expressions, with respect to each other, is person-independent.

1) Signature of an Expression:

As the organization of expressions is similar between subjects, one expression can be uniquely defined by its relative position to others expressions, for instance, the 8 plausible expressions created in subsection 4.C.

By computing a Delaunay tessellation on the first components of the appearance parameters of these 8 expressions plus neutral face, we get a manifold that approximates the appearance space of the subject. Each new expression is projected onto this manifold and we defined the direction-intensity signature of the new expression by:

- The direction is the barycentric coordinates of the projection on the outer surface of the manifold.
- The intensity is the Euclidian distance between the neutral and the expression.

Fig. 11 shows an example of the computation of the direction-intensity signature.

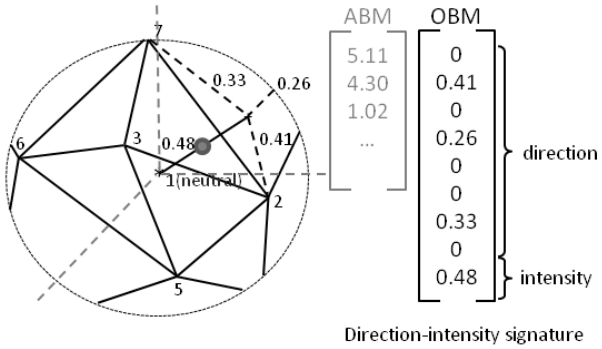


Fig. 11 Transformation from appearance parameters to direction-intensity signature.

2) Person-Independent Expression Space:

As the direction-intensity signature is relative, it is independent of the subject. The expression space is the space of the images defined by these signatures. Fig. 12 shows an extract of a video sequence that displays a smile in this space. The expression space has been unfolded in 2D. Each point corresponds to one image. The direction is given by the position of the point. The intensity is given by the size of the point.

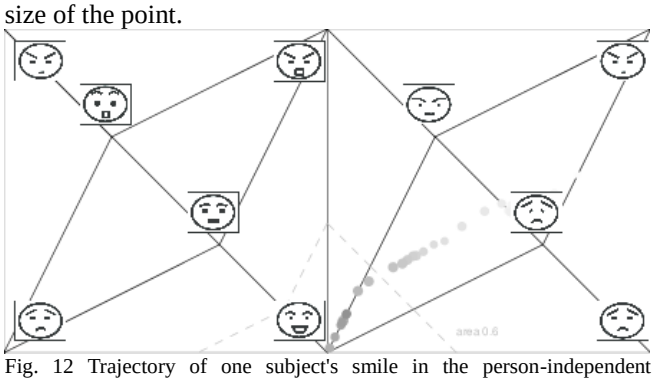


Fig. 12 Trajectory of one subject's smile in the person-independent organized expression space.

E. Facial Expression Extraction

As the expression space is person-independent, the recognition of one expression can be achieved by a basic algorithm. For a given window, we define an area and compute the percentage of frames in this area. The direction of one expression is given by barycentric coordinates of the encompassing triangle and the intensity is between 0 (neutral) and 1 (high intensity). In our system, a smile is defined by a direction that is close to the expression E4 (corresponding to a coefficient above 0.6 - see Fig. 12) and an intensity greater than 0.3. The feature 'laughter' is defined by the percentage of images representing an expression of smile during a time window of 40 seconds.

[1] The database is available at http://www.rennes.supelec.fr/immemo/

TABLE 1
MEAN CORRELATION BETWEEN RELEVANT FEATURES AND EMOTIONAL DIMENSIONS.

Dimensions	Laugh	Body movement	Speaking turns	Speech rate	Response time	Conversation time (Square-wave signal)
Arousal	**0.30**	0.15	-0.02	0.08	**0.43**	0.19
Valence	**0.41**	0.08	0.09	0.03	**0.12**	-0.04
Power	0.10	-0.02	-0.13	0.11	**0.56**	0.26
Unexpectancy	0.11	0.03	**0.25**	-0.03	-0.10	**-0.21**

5. Training Process

This section presents the training process used to select relevant features, that are used in input of both fusion methods (subsection 5.A) and define the fuzzy rules of the fuzzy inference system (subsection 5.B) and the centers of the radial basis function system (subsection 0).

A. Correlation Between Relevant Features and Emotional Dimensions

To find the source of the main variations of the 4 emotional dimensions, we computed the correlation between the ground truth labels of each sequence of development database and a signal that gives one of the relevant features described in section 3. We then compute the average value of these correlation coefficients. A high value of the mean value of the correlation coefficients indicates that the feature can be used to define the global shape of the variations of the emotional dimension. The results are given in TABLE 1. To avoid the impact of the response time of the rater, the first seconds of the sequences have been removed from the computation of the correlation of the features other than response time.

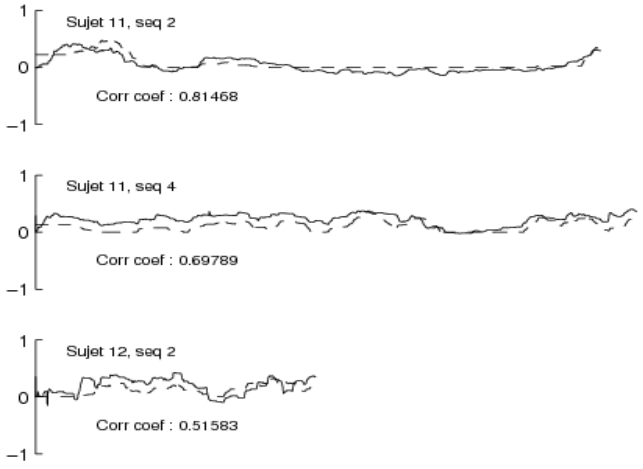


Fig. 13 Correlation between laughter and valence for three sequences of conversations. The solid lines represent the ground truth valence and the dotted lines represent the detection of the variation of the laughter using the proposed feature extraction. The value of the correlation between laughter and valence as well as the subject and sequence of the conversation is displayed for each conversation.

As expected, the duration of high intensity smile gives relevant information on the valence with a mean correlation of 0.41. Fig. 13 illustrates this observation with 3 examples of conversations. The figure shows high correlation between laughter and valence, which is what was expected, since laughing certainly means that the person is positive.

TABLE 1 also show that laughter gives information on arousal with a mean correlation of 0.30. Indeed, when subjects laugh, they are active. Fig. 14 illustrates this observation with the same 3 examples of conversations.

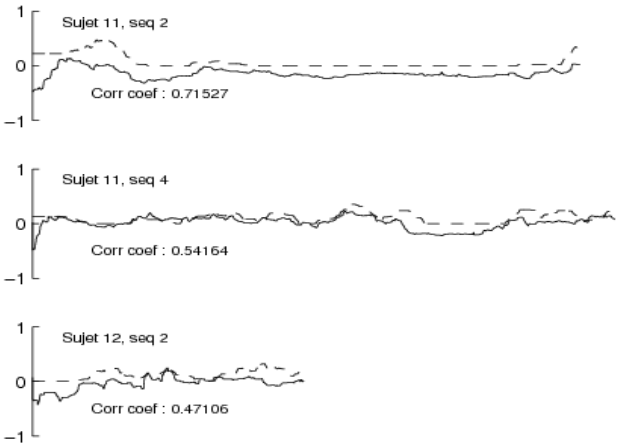


Fig. 14 Correlation between laughter (dotted line) and arousal (solid line) for three sequences of conversations.

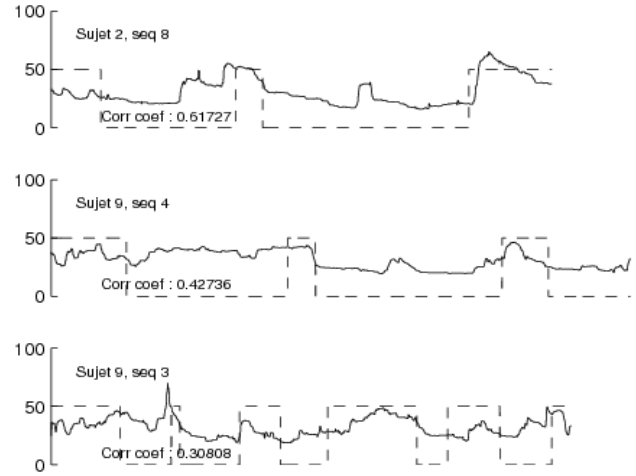


Fig. 15 Correlation between speaking turns (dotted line) and expectancy (solid line) for three sequences of conversations.

TABLE 1 shows that body movement gives information on arousal with a mean correlation of 0.15. Nevertheless, the correlation is low (even lower than laughter), so that we did not considered this value as relevant enough and did not take this value in input of the fusion systems.

The structure of the speaking turns (long or short sentences) gives information on expectancy with a mean correlation of 0.25. Indeed, when subjects speak (long sentences), they are not surprised (as the agent mainly provides backchannels), whereas when they answer and give short answers (short sentences), that may signify that the conversation is unexpected. Fig. 15 illustrates this observation with 3 examples of conversations.

Speech rate seems to be linked with power, but the correlation is low (0.11). This means that sometimes, when subjects speak fast, they are confident. As the correlation is low, we did not considered this value as relevant enough and did not take this value in input of the fusion systems.

The response time of the rater characterizes arousal and power with high correlation of 0.43 and 0.56 respectively. This is due to the fact that there is a big difference between the mean value of the signal and the initial value for these two dimensions as shown in Fig. 5. Fig. 16 illustrates this observation on the power dimension with 3 examples of conversations.

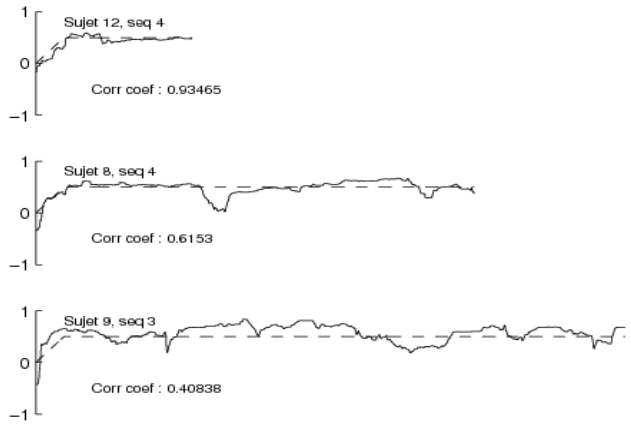


Fig. 16 Correlation between response time of the rater (dotted line) and power (solid line) for three sequences of conversations.

The conversation time (square-wave signal at the beginning of the conversation) confirms the global change in expectancy during a conversation, with a mean correlation of -0.21. Fig. 17 illustrates this observation on the power dimension with 3 examples of conversations. High values of this feature for arousal and power (0.19 and 0.26 respectively) are due to the response time of the rater.

Finally, the impact of the agent's emotional type cannot be measured in terms of correlation, since it gives a constant value over the entire sequence, which is a statistical mean value of valence and arousal. Nevertheless, this statistical mean value is used in the fuzzy rules to define the offset of the sequence (subsection 5.B) and as a component of the input feature vector for the representative samples extraction of the radial basis function system (subsection 0).

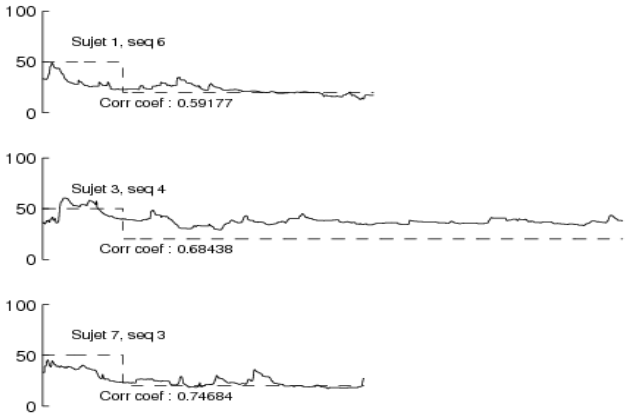


Fig. 17 Correlation between conversation time (dotted line) and expectancy (solid line) for three sequences of conversations.

B. *Rules of the fuzzy inference system*

To fuse the multi-modal features, we used a classical Mamdani type fuzzy inference system [17] where

- the fuzzy operator AND is product,
- the fuzzy implication is product,
- the aggregation is performed by sum,
- defuzzification by centroid method.

The inputs are the relevant features selected according to subsection 5.A. The fuzzy rules of the system are listed in TABLE 2. They are directly derived from the above analysis.

TABLE 2
FUZZY RULES OF THE SYSTEM FOR EACH DIMENSION : VALENCE, AROUSAL, POWER AND EXPECTANCY. RT: RESPONSE TIME OF THE RATER. VL: VERY LOW, L: LOW, AL: AVERAGE LOW, AAL: BETWEEN AL AND A, A: AVERAGE, AH: AVERAGE HIGH, H:HIGH, VH: VERY HIGH

Rules	**Ar.**	**Va.**	**Po.**	**Ex.**
During RT	VL	AAL	VL	
Not RT			A	
Not RT and Agent is Poppy	H	H		
Not RT and Agent is Spike	H	AL		
Not RT and Agent is Obadiah	L	L		
Not RT and Agent is Prudence	A	AH		
Not RT and Agent is Unknown	A	A		
Not RT and Laughter is High	VH	VH		
Sentences are long				VL
Sentences are short				VH
Discourse is beginning				VH
Discourse is established				VL

C. *Radial Basis Function System*

This system is inspired from [19]. The inputs are the relevant features selected according to subsection 5.A, concatenated in an input vector.

TABLE 3
RESULTS OF THE 2 FUSION SYSTEMS : FUZZY INFERENCE SYSTEM AND RADIAL BASIS FUNCTION SYSTEM. MEAN CORRELATION COEFFICIENTS BETWEEN PREDICTION AND GROUND TRUTH. AS COMPARISON, THE LAST ROW SHOWS THE MEAN CORRELATION COEFFICIENT BETWEEN ONE RATER AND THE OTHER ONES, AND THE LAST ROWS OF THE TEST SHOWS THE RESULTS OF THE WINNER AND OF THE THIRD OF THE CHALLENGE AVEC2012.

Dimensions	**Training**		**Development**		**Test**				**Raters**
	FIS	RBF	FIS	RBF	FIS	RBF	[19]	[23]	
Challenge's position					2nd		1rst	3rd	
Arousal	.40	.36	.52	.47	.42	.42	.61	.36	.44
Valence	.39	.40	.47	.43	.42	.42	.34	.22	.53
Power	.61	.59	.59	.58	.57	.57	.56	.48	.51
Expectancy	.37	.37	.30	.32	.33	.32	.31	.33	.33
Mean	.44	.43	.47	.45	.43	.43	.46	.34	.45

The first step is the extraction of representative samples of relevant input features. To perform this task, we use k-means clustering. The centers of the clusters are taken as the representative samples for the emotional dimension. The emotional label associated with each sample is the mean of the labels of the cluster. TABLE 4 shows the 5 representatives samples (5 clusters) computed for the arousal dimension.

TABLE 4
REPRESENTATIVES SAMPLES OBTAINED BY K-MEANS CLUSTERING FOR AROUSAL DIMENSION.

Response time	0.99	0.99	0.99	**0.32**	0.99
Laughter	0.04	0.02	0.03	0.09	**0.25**
Agent's arousal	**-0.24**	**-0.08**	**0.04**	-0.06	-0.00
Arousal	-0.19	-0.11	0.03	-0.28	0.05

The second step is the prediction. The prediction is performed via radial basis functions centred on the representative samples previously computed. Let $\{\mathbf{x_j} \in \mathbb{R}^n, j \in 1,m\}$ be the feature vectors of the m representative samples obtained after the clustering step, and $\{\mathbf{y_j}, j \in 1,m\}$ be the associated labels. The prediction for a sample s described by feature vector $\mathbf{x_s} \in \mathbb{R}^n$ is given by :

$$\hat{y}(s) = \frac{\sum_{j=1}^{m} e^{-\frac{\|\mathbf{x_x}-\mathbf{x_j}\|^2}{\sigma^2}} y_j}{\sum_{j=1}^{m} e^{-\frac{\|\mathbf{x_s}-\mathbf{x_j}\|^2}{\sigma^2}}}$$

where the distance used is the Euclidian distance, and σ is the spread of the radial basis function.

6. Results and Discussion

A. Global Results

TABLE 3 shows the results of both fusion systems on training, development and test databases. The learning has been performed on training and development databases. We also add the results of the winners [20] and the third [23] of the challenge on the test set and the mean correlation coefficient between one rater and the other ones (last row of the table).

First we can note the stability of our results over the different databases, whatever the fusion system used, which means that both methods generalize correctly. Even if the values remain low (average of about 0.44), they are similar to those of the human raters used for ground truth labelling (average of 0.45) and of the winner of the challenge (0.46 on the test set), the other challengers were further behind (0.34 on the test set for the third).

The results of the raters also show we get not as good results on valence. To define the valence, we currently use the duration of high intensity smile. Other information on facial expressions such as the lowering of the eyebrows (AU4) could give us information about a decrease in valence and could improve the results.

The difference on arousal and valence between training and development databases is mainly due to laughter information. We could not find smile information for one subject in the training database; the face is half out of the screen.

B. Raters' (Dis)agreement

The values correlation between the human raters used for ground truth labelling are low (average of about 0.45 – see TABLE 3), which means the human raters often disagree on the variations of the emotion. Fig. 18 illustrates this observation with two examples. On the first one, the two raters agree (correlation of 0.80), whereas on the second one, they strongly disagree (negative correlation of -0.22).

These examples show the difficulty of ground truth labellization in terms of continuous emotional dimension. Such disagreement on the ground truth labels may highly influence the learning of automatic recognition systems.

C. Comparison of the Fusion Techniques

TABLE 3 shows same results for both fusion techniques performed on the same features (FIS and RBF). We can remark that almost the same rules are implemented in both systems.

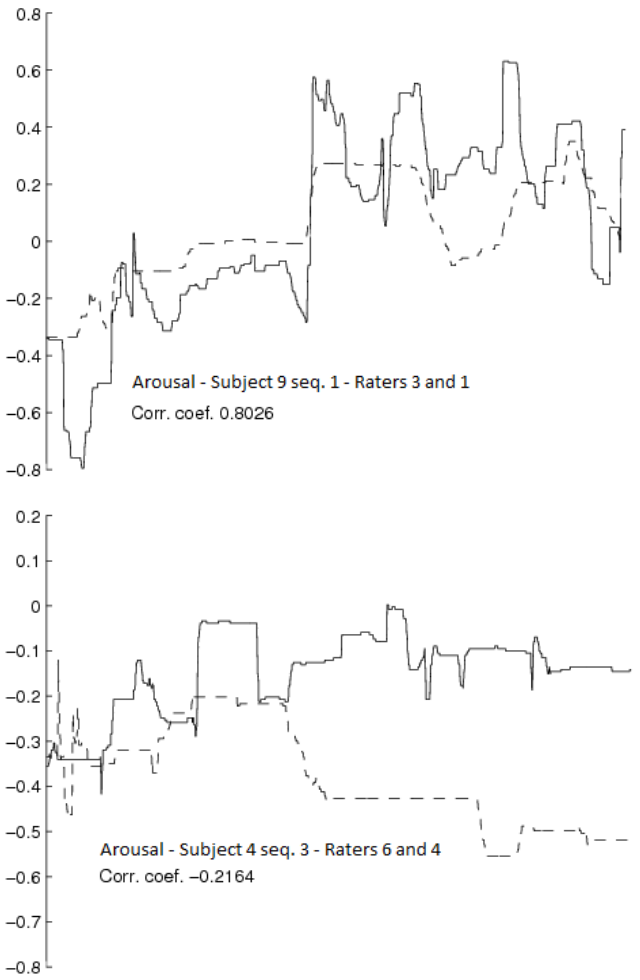


Fig. 18 Comparison of the annotations of two raters. On the top, the two raters agree, on the bottom, they strongly disagree (negative correlation). The emotional dimension as well as the subject and sequence are displayed for each example.

For example, for arousal, we can analyse the clusters of arousal (TABLE 4) by the rules (TABLE 5):

TABLE 5
INTERPRETATION OF THE REPRESENTATIVE SAMPLES INTO RULES FOR AROUSAL DIMENSION.

Cluster	if	arousal is
1	agent is Obadiah	low
2	agent is Prudence or unknown	average
3	agent is Poppy or Spike	high
4	during response time	very low
5	laughter is high	very high

Indeed, in TABLE 4, cluster 4 has a low value for response time (which means that it is during response time), an average value for laughter and agent's arousal and the lowest output value of the clusters for arousal. We interpret these values as 'if we are during response time, whatever the other input values are, the arousal is very low', which is the first rule of TABLE 2. Cluster 5 has a high value for laughter, an average value for agent's arousal, a value near to 1 for response time (which means not during response time) and the highest output value for arousal. We interpret these values as 'if we are not during response time and laughter is high, whatever the other input value is, the arousal is very high', which is the 8th rule of TABLE 2.

In this example, the obtained rules match exactly those of the fuzzy inference system (TABLE 2). This is due to the fact that each input component (response time of the rater, agent's arousal and laughter) has values that dictate an output arousal value.

In the general case, k-means requires more clusters to take into account the combinatorics of the input components. For example, for the valence dimension, the clusters are presented in TABLE 6:

TABLE 6
REPRESENTATIVES SAMPLES OBTAINED BY K-MEANS CLUSTERING FOR VALENCE DIMENSION.

Response time	**.32**	.99	.99	.99	.99	.99	.98
Laughter	.09	**.02**	**.20**	**.01**	**.20**	**.04**	**.27**
Agent's valence	.05	**-.10**	**-.10**	**.10**	**.10**	**.28**	**.28**
Valence	.06	-.10	.20	.08	.21	.25	.33

We can notice that the first cluster corresponds to the first rule of the fuzzy inference system (TABLE 2), but the other clusters combine laughter and agent's valence. For instance, clusters 2 and 3 both correspond to the agent with a low valence (Obadiah), but have a different level of laughter (no laughter for cluster 2 and high laughter for cluster 3). We note the same combination for clusters 4 and 5 (agent with an average valence) and for clusters 6 and 7 (agent with a high valence). In the fuzzy inference system, this combinatorics is directly made by the aggregation, so that the rules do not need to take into account several inputs. We can note the same behaviour with expectancy (see TABLE 7):

TABLE 7
REPRESENTATIVES SAMPLES OBTAINED BY K-MEANS CLUSTERING FOR EXPECTANCY DIMENSION.

Sentences (short/long)	**0**	**0**	**1**	**1**
Discourse (established/beginning)	**0**	**1**	0	1
Expectancy	34.5	38.5	30.6	29.4

Clusters 1 and 2 correspond to the impact of the discourse when the sentences are short: 'When the sentences are short, if the discourse is beginning (cluster 2), the expectancy is very high, whereas if the discourse is established (cluster 1), the expectancy is average (combination of very high and very low)'. Clusters 3 and 4 correspond to the 9th rule 'If sentences are long, expectancy is very low'.

In the general case, the rules are difficult to extract from the analysis of the clusters. On the contrary, the fuzzy inference system uses intelligible rules, facilitates the addition and removal of rules and input data (see discussion 6.D on the context features). But such systems are known to be constrained to few input data to be effective.

As the results are similar for both fusion techniques, we can think that it is not the fusion technique that is a key issue in such systems but the features. This conclusion is

consolidated by the comparison of the results of the proposed radial basis function system and of the AVEC2012 winners' system, that also used k-means clustering and radial basis function prediction (see TABLE 3). Indeed, they obtained quite similar results for power and expectancy (0.56 vs. 0.57 for FIS and RBF for power and 0.31 versus 0.33 for FIS and 0.32 for RBF for expectancy); but significant better results for arousal dimension (0.61 versus 0.42 for FIS and RBF) and worse results for valence dimension (0.34 versus 0.42 for FIS and RBF). This can be analysed by the fact that our global system lacks one or more feature for arousal prediction, and their lacks one or more for valence.

D. Features

As we just said, the features are a key issue in emotion prediction. In this subsection, we discuss the impact of the relevant features.

The impact of smile on arousal can be analysed by the fact that we compute smile with a high intensity on a long duration, that is the characteristics of laughter. Other kinds of smile could be used to improve the results.

The fact that expectancy rises during conversations has to be confirmed by the analysis of other databases displaying conversations to check if this information can be used in a general context. We can note that Ozkan et al. [21] analyzed this observation in a different manner. They explained that the participants perceive the context of the conversation, so that more the participants get engaged to a conversation, the more intense their emotion might get; they consequently used the same time feature for each of the 4 emotional dimensions. In our system, we separate the conversation time used for expectancy (as the participants perceive the context of the conversation, they are less surprised) and the response time of the rater used for the other 3 dimensions. We therefore used two different input signals (see TABLE 1).

TABLE 8
RESULTS OF FIS PROCESSING WITHOUT THE EFFECT OF THE RESPONSE TIME OF THE RATERS (THE FIRST 20 SECONDS HAVE BEEN REMOVED)

Dimension	Train.	Devel.	Test	Raters
Arousal	.22	.32	.21	.39
Valence	.31	.43	.31	.50
Power	---	---	---	.36
Expectancy	.34	.29	.29	.33

As previously said, the ground truth labels of the beginning of the sequences cannot be analysed in term of emotion (due to the response time of the rater), so that we have also computed the results excluding the 20 first seconds of the sequences (TABLE 8). The correlation on power cannot be performed for we only used the response time of the rater for labelling. The results (our results as well as the human raters' ones) are lower especially for arousal and power, for which the response time of the rater plays an important role in the sequence labializations as showed in Fig. 5.

The impact of the emotional agent can be interpreted as empathy and emotion contagion, which may be clue information in general case when audio-visual information is not available or uncertain. In our system, the way of extraction is partly specific to the database (name of the agent), and partly generic (theme of the conversation such as 'fun').

E. Multi-Local AAM vs. Global AAM

In order to compare the performance of the proposed Multi-Model AAM to that of the Global AAM, we plot the Ground Truth Error (GTE) versus the percentage of aligned images for one sequence of the test database. The GTE is the mean of the distance (Euclidean distance) between ground truth (real locations of eyes centers, mouth center and the nose tip) marked manually and these points given by the shape extraction method normalized by the distance between the eyes. The subject in the sequence is smiling most of the time with a smile of varying intensity. Thus the comparison on such a sequence is significant since our system uses a smile detector to detect the emotional dimensions and consequently this smile detector uses AAM results.

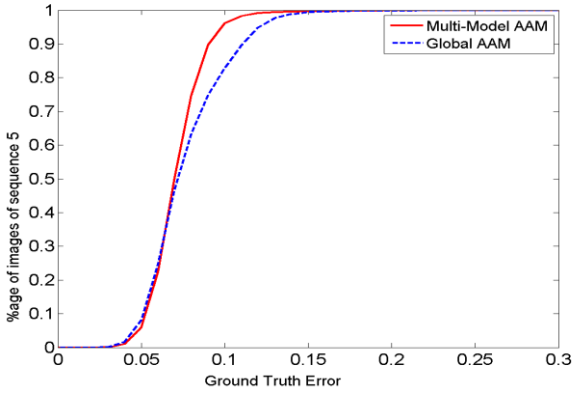


Fig. 19 Comparison between the GTE of the Multi-Model AAM and the Global AAM for one sequence of the tests database.

The GTE of both the MM-AAM and the Global AAM are shown in Fig. 19. The figure shows that with a GTE less than 10% of the distance between the eyes, the MM-AAM is able to extract facial features of 96% of the total images of the sequence, compared to 82% by the Global AAM. Actually for this sequence the local mouth model performs better than global face model at the level of the mouth. So, the MM-AAM chooses the combination of both.

Fig. 20 shows qualitative results on some images of this sequence. This figure shows three cases, in the first case, the subject is smiling wide, in the second, he smiles a small smile after a wide one and, in the third, he opens his mouth while speaking.

As we see, in the first case, the global model fails to give precise results at the level of the mouth because of the wide smile. However the MM-AAM gives the precise result because of its local mouth model. In the second case, the GF-AAM fails because the AAM parameters are initialized by those of the preceding image which is a wide smiling one. In the third, the small contrast between the teeth and

the skin makes a global model fails while a local one does not.

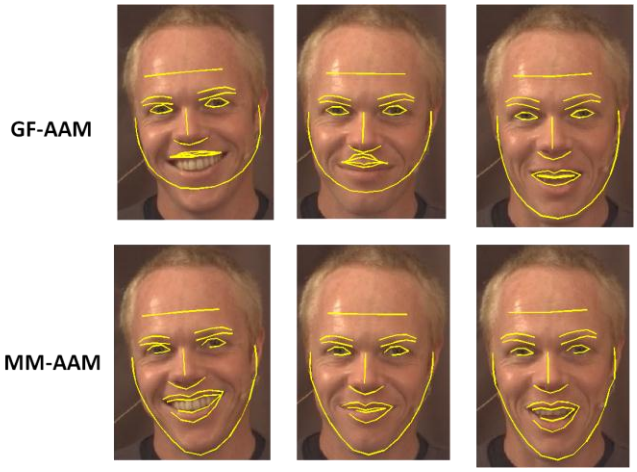


Fig. 20 Comparison between the GF-AAM (top row) and the MM-AAM (bottom row) on one sequence of the test database.

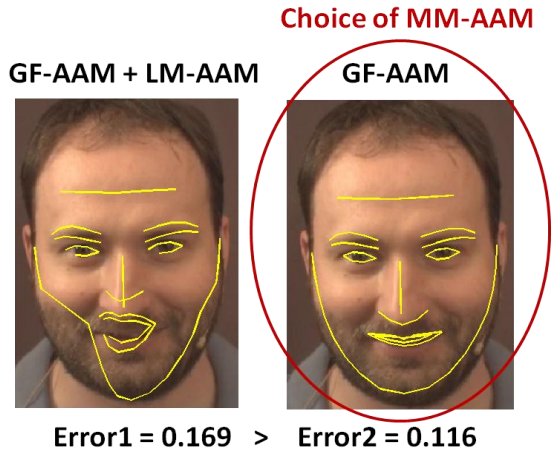


Fig. 21 Example of the MM-AAM in the case where the algorithm chooses the GF-AAM rather than the combination of the GF-AAM and the LM-AAM.

Fig. 21 shows the results of both the GF-AAM and the combination of the GF-AAM and the LM-AAM for another sequence of the test database. In the case of this sequence, the local mouth model performs poorer than the global model. The reason is that the subject has a beard and the local model was not trained on such subjects. The global model makes use of the relationships between the upper part and the lower one to converge even if the training database does not contain such subjects. Thus the MM-AAM chooses the results of the GF-AAM rather than the combination of both for most of the frames of the sequence. As we see from the figure, the GTE curves coincide.

As a conclusion, employing the MM-AAM is efficient in alternating between results of a global AAM and a local one according to the one that performs better which permits to take advantage of both global and local frameworks.

7. Conclusion

This paper has presented a facial expressions space that takes into account the morphology of the subject, and that can effectively and continuously define facial expressions. It is based on the spatial organization of expressions, one with respect to the others. This organization is invariant among the subjects. As the representation is relevant, expression recognition can then be performed with simple algorithms. Here we used an intensity-area detector to extract the high-level feature ‘laughter’.

To analyze the impact of the fusion methods in the global system, this facial expression recognition was integrated into two global methods for the detection of emotion. The same different modalities (audio, video and context) are merged either with a fuzzy inference system or a radial basis function system. They both calculate the 4 emotional dimensions: valence, arousal, power and expectancy. The experiments show that the choice of the fusion technique little impacts the results, which seems to say, that the feature extraction is the key issue of emotion detection.

Contrary to statistical systems, in which learning must be reprocessed for each new database or context, fuzzy inference system can be easily adapted by removing or adding rules that are specific to the database or to the context for real life scenarios.

The results of correlation between ground truth and the obtained values (correlation coefficient of 0.43 on average on the test set) show that there are still improvements to do in order to determine the variations of emotions, even if we perform in average as good as human raters. Adding other kinds of smile and eye brow movements could improve the results.

Acknowledgment

This research has been conducted with the support of Immemo (french ANR project) and Replica (ANR techsan).

Portions of the research in this paper use Semaine Database collected for the Semaine project (www.semaine-db.eu [18] [24]).

We thank Catherine Pelachaud for her advice.

References

[1] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face-Pain expression recognition using active appearance models. *Image and Vision Computing*, 2009.

[2] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, 2004.

[3] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. In *Analysis and Modeling of Faces and Gestures, IEEE International Workshop on*, Nice, France, 2003. IEEE Computer Society.

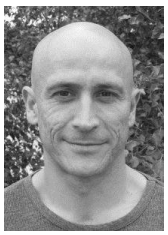
[4] Y. Cheon and D. Kim. Natural facial expression recognition using differential-AAM and manifold learning. *Pattern Recognition*, 2009.
[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2001.
[6] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. Feeltrace : An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, 2000.
[7] P. Ekman and W. V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.
[8] P. Ekman, W. V. Friesen, and P. Ellsworth. *Emotion in the human face*. Cambridge University Press New York, 1982.
[9] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann. Real-time facial expression recognition using a fuzzy emotion model. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, 2007.
[10] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 2003.
[11] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not Two-Dimensional. *Psychological Science*, 2007.
[12] N. Fragopanagos and J. G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 2005.
[13] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. Bimodal fusion of emotional data in an automotive environment. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005.
[14] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. *Artifical Intelligence for Human Computing*, 2007.
[15] G. Langs, P. Peloschek, R. Donner13, and H. Bischof. A clique of active appearance models by minimum description length, 2005.
[16] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *Proceedings of the 9th international conference on Multimodal interfaces*, 2007.
[17] E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 1975.
[18] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, 2010.
[19] S. Moore, R. Bowden, and U. Guildford, The effects of pose on facial expression recognition, *in Proc. British Machine Vision Conf.*, 2009.
[20] J. Nicolle, V. Rapp, K. Bailly, L. Prevost and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012.
[21] D. Ozkan, S. Scherer and L.P. Morency. Step-wise emotion recognition using concatenated-HMM. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012.
[22] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008.
[23] A. Savran, H. Cao, M. Shah, A. Nenkova and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012.
[24] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012 : The continuous Audio/Visual emotion challenge. In to appear in Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012, Santa Monica, 2012.
[25] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011 : The first international Audio/Visual emotion challenge. *Affective Computing and Intelligent Interaction*, 2011.
[26] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion recognition based on joint visual and audio cues. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006.
[27] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011.
[28] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. *Computer Vision in Human-Computer Interaction*, 2005.
[29] C. Soladie, N. Stoiber, and R. Seguier. A new invariant representation of facial expressions : definition and application to blended expressions recognition. In *to appear in Proc. Image Processing (ICIP), 2012 IEEE International Conference on*, 2012.
[30] C. Soladié, H. Salam, C. Pelachaud, N. Stoiber and R. Seguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012
[31] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition-a new approach. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2004.
[32] N. Stoiber, R. Seguier, and G. Breton. Automatic design of a control interface for a synthetic face. In *Proceedings of the 13th international conference on Intelligent user interfaces*, 2009.
[33] Y. L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. *Handbook of face recognition*, 2005.
[34] P.A. Tresadern, H. Bhaskar, SA. Adeshina, C.J. Taylor, and T.F. Cootes. Combining local and global shape models for deformable object matching. *In Proc. British Machine Vision Conference*, 2009.
[35] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011.
[36] Z. Xu, H. Chen, and S.C. Zhu. A high resolution grammatical model for face representation and sketching. *In Computer Vision and Pattern Recognition. IEEE Computer Society Conference on*, 2005.
[37] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, 2011.
[38] L. Zalewski and S. Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. *In Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2005.
[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2009.

[40] C. Zhang and F S Cohen. Component-based active appearance models for face modelling. *In Advances in Biometrics, Springer*, 2005.
[41]

Catherine Soladié is a PhD Student in SCEE (Communication and Electronic Embedded Systems) lab of Supelec. Her research focuses on facial expressions analysis on unknown subjects.

Hanan SALAM received the B.Eng in Electrical and Computer engineering from the Lebanese University, Beyrouth, Lebanon, and the M.Eng in Signal and Image processing from the Ecole Centrale, Nantes, France in 2010. She is currently working to get her PHD degree at the SCEE (Communication and Electronic Embedded Systems) lab of Supélec, Rennes, France. Her research interests include face analysis and eye gaze detection for Human-Machine Interface.

Nicolas Stoiber graduated from the engineering school Supelec in France in 2005. He then obtained a Master of Science in Information and Multimedia Technology at the Technische Universitt Mnchen through a double degree program in 2007. In 2010, he completed a PhD in the field of facial expression analysis and realistic facial animation synthesis. He then joined the founders of company Dynamixyz as an expert on facial movements tracking and modeling. He has since been leading the company R&D work on image analysis, motion capture and animation and human facial expressions modeling.



Renaud Séguier received the PhD degrees in Signal Processing, Image, Radar in 1995 and the HDR (Habilitation Diriger des Recherches) in 2012 from the University of Rennes I. He worked one year in Philips R&D department on numerical TV and Mpeg2 transport-stream. He joined SCEE (Communication and Electronic Embedded Systems) lab of Suplec in 1997 since when he is Assistant Professor and now Professor in Image Processing, Artificial Life and Numerical Implementation. His current research focuses on face analysis and synthesis for object video-compression and Human-Machine Interface.