A Multimodal Fuzzy Inference System using a Continuous Facial Expression Representation for Emotion Detection

Catherine Soladié
SUPELEC/IETR, Team SCEE
Avenue de la Boulaie
35576 Cesson-Sévigné
France
catherine.soladie@supelec.fr

Hanan Salam
SUPELEC/IETR, Team SCEE
Avenue de la Boulaie
35576 Cesson-Sévigné
France
hanan.salam@supelec.fr

Catherine Pelachaud
CNRS - Telecom ParisTech
37/39 rue Dareau
75014 Paris France
catherine.pelachaud@telecom-paristech.fr

Nicolas Stoiber
Dynamixyz
80 avenue des Buttes de Coesmes
35700 Rennes France
nicolas.stoiber@dynamixyz.com

Renaud Séguier
SUPELEC/IETR, Team SCEE
Avenue de la Boulaie
35576 Cesson-Sévigné
France
renaud.seguier@supelec.fr

ABSTRACT

This paper presents a multimodal fuzzy inference system for emotion detection. The system extracts and merges visual, acoustic and context relevant features. The experiments have been performed as part of the AVEC 2012 challenge. Facial expressions play an important role in emotion detection. However, having an automatic system to detect facial emotional expressions on unknown subjects is still a challenging problem. Here, we propose a method that adapts to the morphology of the subject and that is based on an invariant representation of facial expressions. Our method relies on 8 key expressions of emotions of the subject. In our system, each image of a video sequence is defined by its relative position to these 8 expressions. These 8 expressions are synthesized for each subject from plausible distortions learnt on other subjects and transferred on the neutral face of the subject. Expression recognition in a video sequence is performed in this space with a basic intensity-area detector. The emotion is described in the 4 dimensions : valence, arousal, power and expectancy. The results show that smile is the expression that is the most meaningful for valence detection and can also be used to improve arousal detection. The main variations in power and expectancy are given by context data.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Correlation and regression analysis*; I.4.10 [**Image Processing and Computer Vision**]: Image Representation; I.5.1 [**Pattern Recognition**]: Models

General Terms

Algorithm, Performance

Keywords

Facial Expression Representation, Transfer learning, Fusion Techniques, Fuzzy Inference System, Context in Emotion Recognition

1. INTRODUCTION

Our aim is to develop a method to help elderly to age in place, which is a key goal of many new government programs. One solution is to be able to detect when changes in the elderly's emotional state occur and to send an alarm if necessary. Such systems have to be non intrusive, easy to use and agreeable. This paper explores a system based on a camera which can be included into a set-top box. 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. Such a system requires :

- accuracy to detect changes in emotion;
- exhaustivity to have continuous information of the emotional state;
- robustness to cope with the diversity of the subjects;
- flexibility to adapt itself to a subject without a previous learning phase.

In this paper, we focus on a continuous representation of facial expressions, that enables recognizing emotions on unknown subjects. Facial expressions are relevant visual features for emotion detection, especially the smile. The experiments have been performed as part of the AVEC 2012 challenge [18], which is done over a collection of audio-video sequences, displaying conversations between an emotional agent and an unknown subject. The emotion is described in the 4 dimensions : valence, arousal, power and expectancy [11]. To enhance our results for the challenge, we added acoustic and context features to facial expression features. Context is information specific to Semaine database [16], for example which is the emotional agent that discusses with the

subject. The relevant features from audio, video and context are used in a multimodal fuzzy inference system.

Many expressions recognition systems have been proposed in the last decade. However, they do not answer all the cited constraints. The choice of representation is known to influence the recognition performance. Before presenting our system, we will briefly present a state-of-the-art on multimodal emotion recognition systems and on facial expression recognition.

1.1 Multimodal emotion recognition

Multimodal emotion recognition has developed in recent years [29]. 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 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 [24], neural network methods [12, 14] or Bayesian network methods [20]. 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 [17].

The majority of the systems focus on the classification of discrete emotions [2, 13, 20]. Some systems evolved towards a dimensional representation of emotions (activation, valence, evaluation) but the output value remains discrete [14]. Most systems merge data from the two modalities : acoustic (prosody) and visual (facial expressions). Speaking turns or the agent's emotional type are rarely taken into account. Multimodal automatic emotion analysis is still a challenging problem as the last Audio/Visual Emotion Challenge (AVEC 2011 [19]) shows. The results of the challenge were not relevant and struggled to perform better than chance.

1.2 Facial expression recognition

Many expressions recognition systems have been proposed in the last decade [10, 26]. 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 [15, 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 appearance features (shape and/or texture features). Other systems extract Action Units defined in the FACS system [7]. In the last few years, some continuous representations of facial expression have been proposed, using manifold learning [3, 22, 4, 25]. 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 [27]) consisted of the recognition of discrete emotion won by [28] and detection of AUs that our team won [21] with the ISIR laboratory. Even if the results were encouraging, the recognition rates remained low, especially for AU detection.

1.3 Overview of our approach

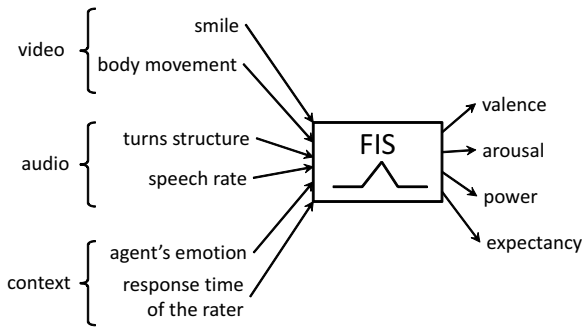

Figure 1: Overall view of the proposed method : a fuzzy inference system transforms the relevant features from video, audio and context into 4 emotional dimensions.

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 subject. The article focuses on facial expressions detection. The main contribution of the article is 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. Another main contribution is the continuous representation of facial expressions, 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.

The remainder of this paper is organized as follows. In section 2, we describe the global process for multimodal emotion recognition and we define the relevant features extraction. Section 3 focuses on the facial expression extraction. Section 4 presents the training process performed on training and development databases and the resulting fuzzy rules. Section 5 shows and discusses the results of the challenge on the test database. Section 6 concludes the paper.

2. A MULTIMODAL FUZZY INFERENCE SYSTEM

This section presents the global system for emotion detection. The system is based on a fuzzy inference system (see subsection 2.1). The relevant features extraction is described in subsection 2.2.

2.1 Global process

The overall process is presented in figure 1. The system is based on a fuzzy inference system that takes in input the relevant features that result from emotional states. The fuzzy inference rules are defined from the analysis of the data of the training and development databases (section 4). The output information is defuzzified to output the 4 emotional dimensions, namely valence, arousal, power and expectancy.

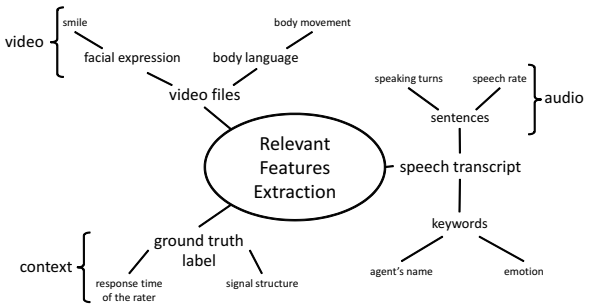

Figure 2: Sources of the relevant features : video files, speech transcripts and emotional labels.

2.2 Relevant features extraction

The choice of the features to extract has been done according to the state-of-the-art on features that may help identifying an emotion and also on 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. The features extraction is made from 3 different data sources : videos, speech transcripts and ground truth labels (see figure 2).

2.2.1 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 3).

Concerning the body language, we computed the global movement of the head pose in the scene. The video data are analyzed 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.

2.2.2 From speech transcripts

The features extracted from audio mode come from the analysis of the sentences and keywords.

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), 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. Figure 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 behaviors of high valence and high arousal. To find automatically which is the emotional agent of the sequences, 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.

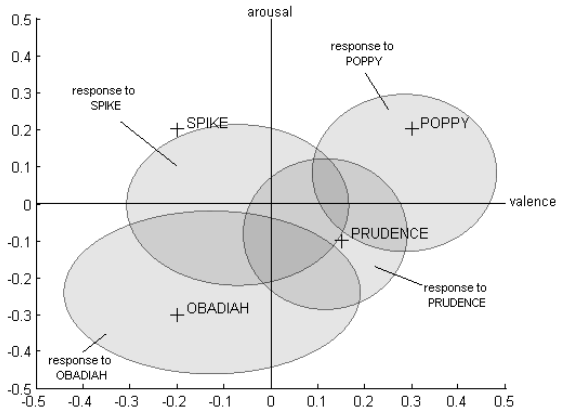

Figure 3: Displayed emotional response of subjects when interacting with agent. The emotion of the agent of SEMAINE is showed by a cross (Spike is aggressive; Poppy is cheerful; Obadiah is gloomy; and Prudence is pragmatic). The ground truth labels of the subjects responding to these agents are in the ellipses.

2.2.3 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 figure 4) for arousal and power. This may be due to the initialization of the tool used to rate and to the response time of the rater. We modeled this behavior

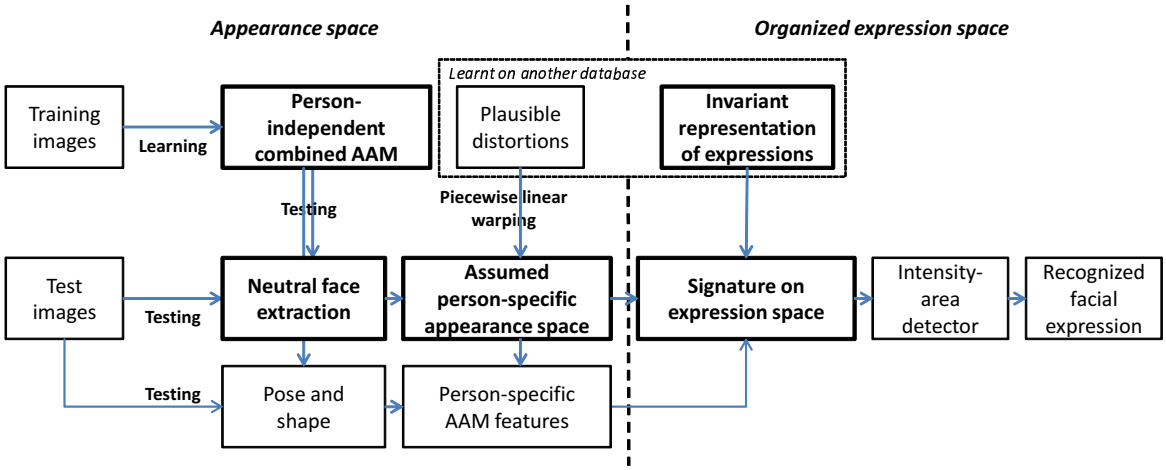

Figure 5: Overall view of the facial expression extraction. The neutral face of each subject and the shape of all the images are extracted using a person-independent combined AAM. An assumed person-specific appearance space is created by applying plausible distortions on the neutral face of the given subject. The person-specific appearance space is transformed into the expression space using an invariant organization of expressions.

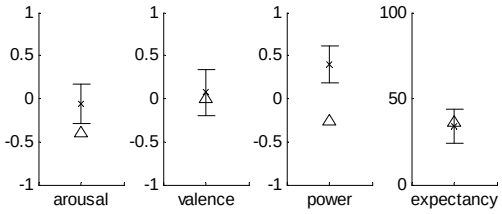

Figure 4: Impact of the response time of the raters on the ground truth labels : the triangle shows the value of the label at the beginning of the sequence and the cross shows the mean and standard deviation of the labels, for each dimension.

with a feature as a decreasing linear function on the first 20 seconds of the conversation.

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

3. FACIAL EXPRESSION EXTRACTION

This section presents the main contribution of the article. After a brief description of the overall process (section 3.1), each main step of the facial expression extraction is described (sections 3.2, 3.3, 3.4, 3.5).

3.1 Overview of the process

The global overview of the process is presented in figure 5. 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

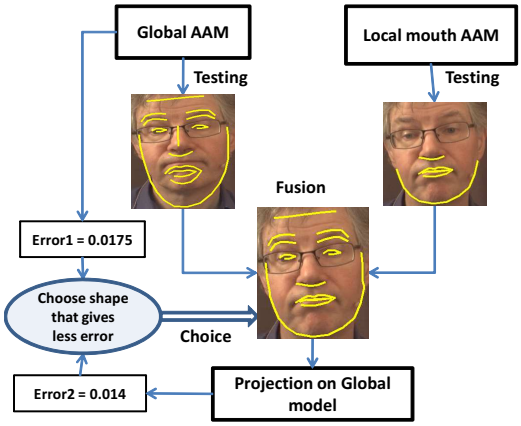

Figure 6: Example of person-independent Multi-Model AAM (MM-AAM).

by a person-independent AAM (section 3.2). The second step computes a person-specific appearance space (section 3.3). The third step transforms this appearance space into a person-independent expression space (section 3.4). The last step performs the expression recognition (section 3.5).

3.2 Multi-Model 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, we use a 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

Figure 7: 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.

the Multi-Model AAM (MM-AAM) (cf. figure 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 $\bar{s}$ of GF-AAM;
(b) Find the shape parameters b_s of S_{MM} using $s = \bar{s} + V_s b_s$. V_s are 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 are the texture eigenvectors of the GF-AAM;
(e) Concatenate b_s and b_g: $b = \binom{W_s b_s}{b_g}$. $W_s$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.

3.3 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.

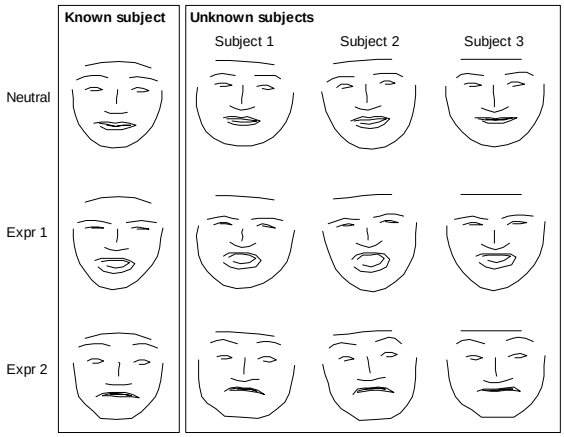

Figure 8: The plausible distortions are learnt from a known subject and applied on the neutral face of the unknown subjects. (Neutral) Neutral face of the known subject and of 3 unknown subjects. (Expr x) Real expression of the known subject and plausible expression of the unknown subjects.

3.3.1 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. Figure 7 shows some examples of neutral faces.

3.3.2 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 figure 8). 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.

3.4 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 [23] that the organization of 8 expressions, with respect to each other, is person-independent.

Instead of describing one expression by its appearance features (which means taking into account the morphology of

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

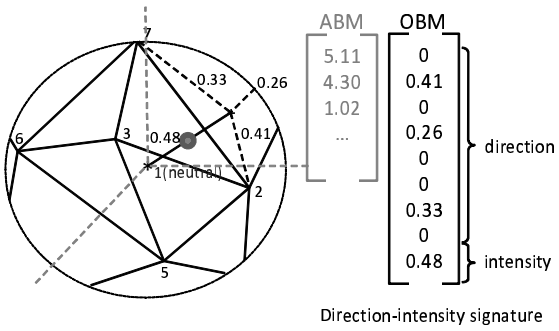

Figure 9: Transformation from appearance parameters to direction-intensity signature.

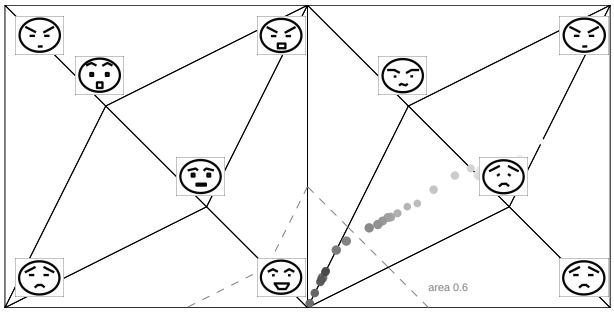

Figure 10: Trajectory of one subject's smile in the person-independent organized expression space.

the subjects), we describe one expression by its relative position to others. We previously showed [23] that the organization of 8 expressions, with respect to each other, is person-independent.

3.4.1 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 3.3.

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 norm of the vector neutral-expression.

Figure 9 shows an example of the computation of the direction-intensity signature.

3.4.2 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. Figure 10 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.

3.5 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) and an intensity greater than 0.3. The feature 'smile' is defined by the percentage of images representing an expression of smile during a time window of 40 seconds.

4. TRAINING PROCESS

This section presents the training process used to define the fuzzy rules of the global multimodal process.

4.1 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 2.2. 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.

As expected, smile gives relevant information on the valence. We can notice that it also gives information on arousal. Indeed, when subjects smile, they are active. Body movement gives information on arousal. Nevertheless, the correlation is low, even lower than smile. The structure of the speaking turns (long or short sentences) gives information on expectancy. 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. Speech rate seems to be linked with power, but the correlation is low. This means that sometimes, when subjects speak fast, they are confident. The response time of the rater characterizes arousal and power with high correlation. 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 figure 4. The square-wave signal at the beginning of the conversation confirms the global change in expectancy during a conversation. High values of this feature for arousal and power 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 (figure 3), which is a statistical mean value of valence and arousal. Nevertheless, this statistical mean value

Table 1: Mean correlation between relevant features and emotions dimensions.

Dimensions	Smile	Body movement	Speaking turns	Speech rate	Response time	Square-wave signal
Arousal	**0.3032**	**0.1495**	-0.0207	0.0799	**0.4312**	**0.3178**
Valence	**0.3293**	0.0849	0.0924	0.0301	0.1235	0.0005
Power	0.1003	-0.0200	-0.1250	0.1084	**0.5566**	**0.3923**
Unexpectancy	0.0410	0.0324	**0.2535**	-0.0263	-0.0979	**-0.2220**

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 Smile is high	VH	VH		
Sentences are long				VL
Sentences are short				VH
Discourse is beginning				VH
Discourse is established				VL

Table 3: Results (mean correlation coefficients) on training, development and test databases. As comparison, the last row shows the mean correlation coefficient between one rater and the other ones (the raters are those of SEMAINE used for ground truth labellization [16]).

Dimension	Train.	Devel.	Test	Raters
Arousal	0.3967	0.5180	0.4202	0.4421
Valence	0.3894	0.4705	0.4187	0.5175
Power	0.6140	0.5856	0.5710	0.5110
Expectancy	0.3684	0.2997	0.3275	0.2811
Mean	0.4421	0.4685	0.4343	0.4379

is used in the fuzzy rules to define the offset of the sequence (table 2).

4.2 Rules of the fuzzy system

We defined the fuzzy rules of the system. They are listed in table 2.

5. RESULTS

Figure 3 shows the results on training, development and test databases. The values of correlation are provided for each dimension. The mean value over dimensions is also showed. The learning has been performed on training and development databases. First we can note the stability of our results over the different databases, which means that our method generalizes correctly. The difference on arousal between training and development databases is mainly due to smile information. We could not find smile information for one subject in the training database; the face is half out of the screen. We can also notice that the dimensions we best approximate are those for which context information is relevant (arousal and power). Nevertheless, the values remain low (average of about 0.4), meaning there is still work to be done to be able to continuously compute the variation of emotions of unknown subjects.

Figure 3 also shows the correlation coefficients of the raters used for ground truth labellization. We notice that these values are low as well, which means the raters often disagree on the variations of the emotion, especially for expectancy. We perform as good as the raters on arousal, power and expectancy, but less on valence (correlation of 0.4 with our method versus 0.5 for the raters).

6. CONCLUSIONS

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.

This facial expression recognition was integrated into a global method for the detection of emotion. A fuzzy inference system is used to merge different modalities (audio, video and context) and calculate 4 emotional dimensions. Few characteristics by modality were used, when trying to describe the overall trend of the signal. For facial expressions, we only use the smile to compute valence and arousal.

The results of correlation between ground truth and the obtained values (correlation coefficient of 0.4 on average) show that there are still improvements to do in order to determine the variations of emotions. The results of the raters show we get not as good results on valence. To define the valence, we currently use smile, which is a clue for high valence. 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.

7. ACKNOWLEDGMENTS

This research has been conducted with the support of Immemo (french ANR project) and Immersivite (Britanny Region PME project). Portions of the research in this paper use Semaine Database collected for the Semaine project (`www.semaine-db.eu` [16, 19]).

8. 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. 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. *Computer Vision in Human-Computer Interaction*, 2005.

[23] 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.

[24] 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.

[25] 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.

[26] Y. L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. *Handbook of face recognition*, 2005.

[27] 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.

[28] 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.

[29] 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.