Audio-visual emotion recognition: A dynamic, multimodal approach

Jérémie Nicolle, Vincent Rapp, Kevin Bailly, Lionel Prevost, Mohamed Chetouani

Audio-visual emotion recognition: A dynamic, multimodal approach

Jeremie Nicolle[1] Vincent Rapp[1] Kevin Bailly[1] Lionel Prevost[2] Mohamed Chetouani[1]

[1]Univ. Pierre et Marie Curie ISIR (UMR7222)
F-75005 Paris, France
{nicolle, bailly, rapp, chetouani}@isir.upmc.fr

[2]Univ. des Antilles LAMIA
97159 Pointe-à-Pitre, Guadeloupe
lionel.prevost@univ-ag.fr

ABSTRACT
Designing systems able to interact with students in a natural manner is a complex and far from solved problem. A key aspect of natural interaction is the ability to understand and appropriately respond to human emotions. This paper details our response to the continuous Audio/Visual Emotion Challenge (AVEC'12) whose goal is to predict four affective signals describing human emotions. The proposed method uses Fourier spectra to extract multi-scale dynamic descriptions of signals characterizing face appearance, head movements and voice. We perform a kernel regression with very few representative samples selected via a supervised weighted-distance-based clustering, that leads to a high generalization power. We also propose a particularly fast regressor-level fusion framework to merge systems based on different modalities. Experiments have proven the efficiency of each key point of the proposed method and our results on challenge data were the highest among 10 international research teams.

Key-words
Affective computing, Dynamic features, Multimodal fusion, Feature selection, Facial expressions.

ACM Classification Keywords
H5.2 [User Interfaces] User-centered design

I.5.4 [Application] Computer Vision

INTRODUCTION

Intelligent Tutoring Systems (ITS) are computer systems that aim to provide instruction and feedback to learners, (sometime with or) without the intervention of a human expert. Many years after the introduction of ITS in education and professional settings, they have demonstrated their capabilities and limitations. One of these latter is shared with traditional learning: the requirement of individualized learning [20]. Current ITSs are not adaptive due to their lack of interactivity and emotionality. Recent research in Learning Analytics can improve interactivity by predicting one student's learning style [3]. Here, we'll focus on user emotions understanding that belongs to Affective Computing research area [19]. It is well known that emotion can either enhance or inhibit learning [7]. Positive emotions, usually considered as pleasant states, impact positively learning, curiosity and creativity. Kort et al. [11] proposed a four-quadrant learning model where the first dimension is the affect "sign" (positive or negative) and the second one, the "learning activity" (from unlearning to constructive learning). One may notice that this model is not far from Rusell's two dimensional (valence-arousal) model of affect.

So, in ITS and, more generally, in HCI, a current challenge is to give the computer the ability to interact naturally with the user with some kind of emotional intelligence. Interactive systems should be able to perceive pain, stress or inattention and to adapt and respond to these affective states. An essential step towards this goal is the acquisition, interpretation and integration of human affective state within the HCI. To recognize affective states, human-centered interfaces should interpret various social cues from both audio and video modalities, mainly linguistic messages, prosody, body language, eye contact and facial expressions.

Automatic recognition of human emotions from audio and video modalities has been an active field of research over the last decade. Most of the proposed systems have focused on the recognition of acted or prototypal emotions recorded in a constrained environment and leading to high recognition rates. These systems usually describe affects via a prototypal modeling approach using the six basic emotions introduced in the early 70s by Ekman [4]. Another standard way to describe facial expressions is to analyze the set of muscles movements produced by a subject. These movements are called facial Action Units (AUs) and the corresponding code is the Facial Action Coding System (FACS) [5]. The first challenge on Facial Expression Recognition and Analysis (FERA'11) focused on these two kinds of affect description. Meta-analysis of challenge results are summarized in [29]. These methods generally use discrete systems whether based on static descriptors (geometrical or appearance features) and on static classifiers such as Support Vector Machines [27].

However, these descriptions do not reflect real-life interactions and the resulted systems can be irrelevant to an everyday interaction where people may display subtle and complex affective states. To take this complexity into account, this classical description via prototypal modeling approach has recently evolved to a dimensional approach where emotions are described continuously within an affect space. The choice of the dimensions of this space remains an open question but Fontaine [6] showed that four dimensions cover the majority of affective variability: Valence (positivity or negativity), Arousal (activity), Expectancy (anticipation) and Power (control). The Affective Computing research community has recently focused on the area of dimensional emotion prediction and the first workshop on this topic (EmoSPACE'11 [9]) was organized three years ago, followed by the Audio/Visual Emotion Challenge (AVEC'11 [25]).

In this paper, we report the method we proposed to participate to the second edition of AVEC in 2012. Next section will be devoted to a state of art on multimodal affect recognition systems. Based on this latter, we'll describe the system we designed. Then we'll detail consecutively the feature extraction process, the dimensional predictors' training and the final combination. The following sections are dedicated to evaluation and meta-analysis of the challengers. Finally, conclusion and future works are presented.

MULTIMODAL AFFECT RECOGNITION

Usually, the most important parts of multimodal emotion recognition systems are the learning database, the extracted features, the predictor and the fusion method. More precisely, one of the main key points concerns the features' semantic level. Some methods use low-level features. For example, Wollmer et al. [30] propose an approach using features based on the optical flow. Dahmane et al. [2] use Gabor filter energies to compute their visual features. Ramirez et al. [21], conversely, prefer to extract high-level features such as gaze direction, head tilt or smile intensity. Similarly, Gunes et al. [8] focus on spontaneous head movements.

Another key aspect of this new dimensional approach is the need for the system to take the dynamic of human emotions into account. Some methods propose to directly encode dynamic information in the features. For example, Jiang et al. [10] extend the purely spatial representation LPQ to a dynamic texture descriptor called Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP). Cruz et al. [1] propose an approach that aligns the faces with Avatar Image Registration, and subsequently compute LPQ features. Mcduff et al. [12] predict valence using facial Action Unit spectrograms as features. In this study, we focus on mid-level dynamic features, extracted using different visual cues: head movements, face deformations and also global and local face appearance variations. Most methods use visual cues directly as features. In our method, dynamic information is included by computing the log-magnitude Fourier spectra of the temporal signals that describe the evolution of the previously introduced visual cues. Since an accurate and robust system should take advantage of interpreting signals from various modalities, we also include audio features to bring complementary information.

For the prediction step, different machine learning algorithms can be used. Several methods are based on context-dependent frameworks. For example, Meng et al. [14] propose a system based on Hidden Markov Models. Wollmer et al. [30] investigate a more advanced technique based on context modeling using Long Short-Term Memory neural networks. These systems provide the advantage to encode dynamics within the learning algorithm. Another solution is to base the system on a static predictor as, for instance, the well-known Support Vector Machine [1, 24]. Dynamic information being already included in our features, we chose a static predictor. The proposed method uses a kernel regressor based on the Nadaraya-Watson estimator [15]. For selecting representative samples, we perform a clustering step in a space of preselected relevant features.

To merge all visual and vocal information, various fusion strategies may be relevant. Feature-level fusion (also called early fusion) can be performed by merging extracted features from each modality into one cumulative structure and feeding it to a single classifier. This technique is appropriate for synchronized modalities but some issues may appear for unsynchronized or heterogeneous features.

Another solution is decision-level fusion (or late fusion); each extracted feature set feeds one classifier and all the classifier outputs are merged to provide the final response. For example, Nicolaou et al [17] propose an output-associative fusion framework. In our case, the fusion is based on a simple method linearly combining outputs corresponding to the predictions of the four dimensions with different systems to make the final predictions. This way, the system is able to capture the correlations between the different dimensions.

DESIGNED SYSTEM

This is our response to the continuous Audio/ Visual Emotion Challenge (AVEC'12) [26]. This challenge uses the SEMAINE [13] corpus as benchmarking database. This database has been continuously annotated by humans in real-time and a delay between the affect events and the labels has thus been introduced.

The main contributions presented in this paper for affective signals prediction are the followings.

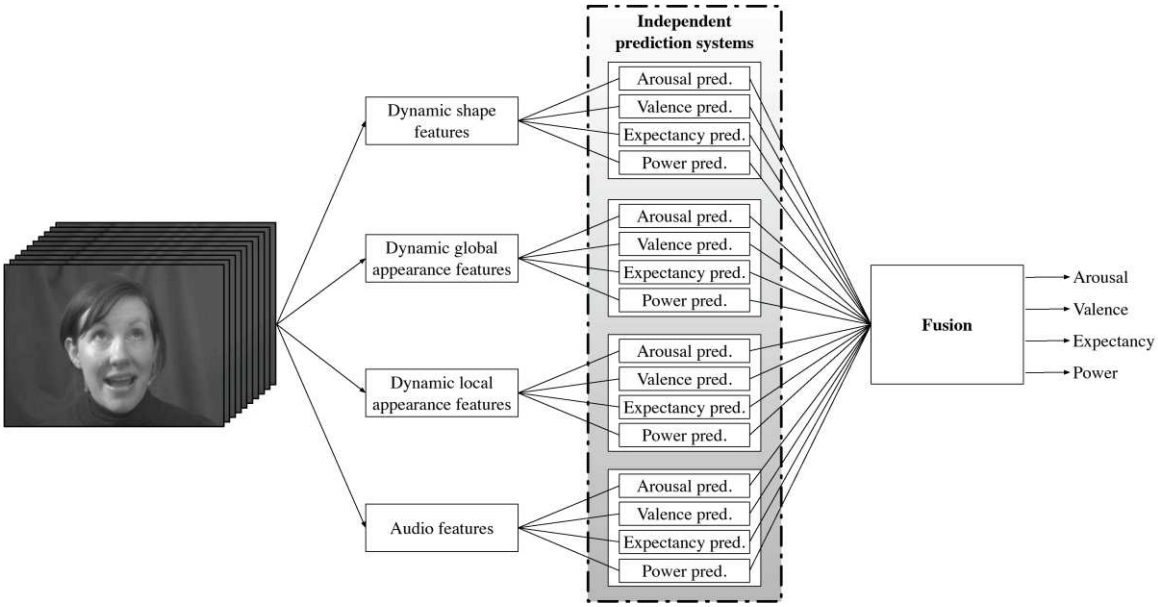

Figure 1: Proposed framework

- The use of the log-magnitude Fourier spectrum to include dynamic information for human emotions prediction.
- A new correlation-based measure for the feature selection process that increases robustness to possibly time-delayed labels.
- A fast efficient framework for regression and fusion designed for real-time implementation.

The proposed framework, presented in Fig. 1 is based on audio-visual dynamic information detailed in the next section. As visual cues, we propose a set of features based on facial shape deformations, and two sets respectively based on global and local face appearance. For each visual cue, we obtain a set of temporal signals and encode their dynamic using log-magnitude Fourier spectra. Audio information is added using the provided audio features. Regarding the prediction, we propose a method based on independent systems for each set of features and for each dimension. For each system, a new correlation-based feature selection is performed using a delay probability estimator. This process is particularly well-adapted to unsure and possibly time-delayed labels. The prediction is then done by a non-parametric regression using representative samples selected via a k-means clustering process. We finally linearly combine the 16 outputs during a fusion process to take into account dependencies between each modality and each affective dimension.

FEATURES

In this section, we present the four different sets of features we used. We propose three multi-scale dynamic feature sets based on video; the fourth one is based on audio.

For the sets of visual cues, we first extract temporal signals describing the evolution of facial shape and appearance movements before calculating multi-scale dynamic features on these signals. The feature extraction process is described in Fig. 2.

Signal extraction

We extract three kinds of signals: one based on shape parameters, and two others based on global and local face appearance.

Shape parameters

The first set of features we used is based on face deformation shape parameters. The initial step of this feature extraction process is face detection performed by Viola and Jones' state-of-art algorithm. Then, we use the 3D face tracker proposed in [22]. It detects the face area and estimates the relative position of 66 landmarks using a Point Distribution Model (PDM). The position of the i^{th} landmark $\mathbf{s_i}$ in the image can be expressed as:

$$\mathrm{s_i}(\mathrm{p}) = s\mathbf{R}(\bar{\mathrm{s}}_\mathrm{i} + \Phi_\mathrm{i}\mathrm{q}) + \mathrm{t}$$

where the mean location of each landmark and the principal subspace matrix are computed from training shape samples using principal component analysis (PCA). Here, $\mathbf{p} = \{s, \mathbf{R}, \mathbf{t}\}$ denotes the PDM parameters, which consist of global scaling s, rotation $\mathbf{R}$ and translation $\mathbf{t}$. Vector $\mathbf{q}$ represents the deformation parameters that describe the deformation of $\mathbf{s_i}$ along each principal direction.

As output of this system, we obtain temporal signals: some of them correspond to the external parameters and give information on the head position, and the others characterize deformations related to facial expressions.

Global appearance
The second set of features we used is based on global face appearance. First, we warp the faces into a mean model using the point locations obtained with the face tracker. This way, the global appearance will be less sensitive to shape variations and head movements, already encoded in the first set. Then, we select the most important modes of appearance variations using PCA. We obtain a set of temporal signals by projecting the warped images on the principal modes.

Local appearance
The third set is based on local face appearance. First, we extract local patches of possibly interesting areas regarding deformations related to facial expressions. We extract an area around the mouth in order to capture smiles, areas around the eyes to capture the gaze direction, around the eyebrows to capture their movements, and areas where the most common expression-related lines can appear (periorbital lines, glabellar lines, nasolabial folds and smile lines). We chose to avoid the worry lines area because of the high probability it has to be occulted by hairs. Then, we use PCA as for the global warped images to compute temporal signals corresponding to the evolution of the local appearance of the face during time.

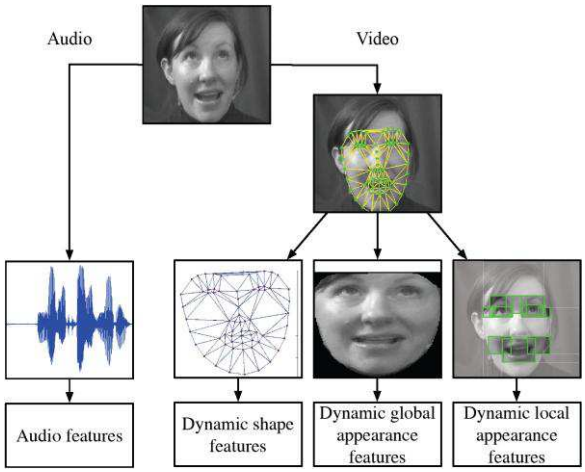

Figure 2: Feature extraction overview

Dynamic features
For each of these three sets, we calculate the log-magnitude Fourier spectra of the associated temporal signals in order to include dynamic information. We also calculate the mean, the standard deviation, the global energy, and the first and second-order spectral moments. We chose to compute these features every second for different sizes of windows (from one to four seconds). This multi-scale extraction gives information about short-term and longer-term dynamics.

Audio features
The last set of features we used is the audio feature set given to the participants of the AVEC'12 Challenge. It contains the most commonly used audio features for the aimed task of predicting emotions from speech (energy, spectral and voice-related features).

Feature normalization
Within a set of features, the range of values can be highly different from one feature to another. In order to give the same prior to each feature, we need to normalize them. A global standardization on the whole database would be a solution but we chose to standardize each feature by subject in order to reduce the inter-subject variability. This method should be efficient under the hypothesis that the amount of data for each subject is sufficiently representative of the whole emotion space.

PREDICTION SYSTEM
Using each of the four feature sets, we make separate predictions for the four dimensions, leading to a total of 16 signals.

Delay probability estimation
The SEMAINE database has been continuously annotated by humans. Therefore, a delay exists between videos and labels, which may significantly corrupt the learning system [16]. We introduce in this paragraph a delay probability estimation method to avoid this issue. Let $y(t)$ be the label signal and $f_i(t)$, $i = \{1, \ldots, n\}$ a set of n features. Making the assumption that the features that are relevant for our prediction will be more correlated to the undelayed label, we can use the sum of the correlations between the features and the τ seconds delayed label signal as a probability index for the label to be delayed by τ seconds. Thus, we can estimate the delay probability $P(\tau)$ as follows:

$$P(\tau) = \frac{1}{A}\sum_{i=1}^{n} r(f_i(t), y(t-\tau))$$

where r is the Pearson correlation coefficient.

$$r(X,Y) = \frac{E(X-\bar{X})E(Y-\bar{Y})}{\sigma_X \sigma_Y}$$

A is the normalization coefficient defined as:

$$A = \int_{-\infty}^{\infty} \sum_{i=1}^{n} r(f_i(t), y(t-\tau))d\tau$$

We calculate $P(\tau)$ for τ varying in a range $[0, T]$ where T is the largest expected delay that we fixed at 20 seconds to obtain an estimate of the delay probability distribution in this range. In our case, the data contain different video sequences. We thus estimate the delay probability as the mean of the delay probabilities estimated for the different sequences. To simplify notations, we refer to this estimate as $P(\tau)$.

In Fig. 3, we represent the four different delay probability distributions that have been learned on the training database for the first feature set. By looking at those distributions' maxima, we identify an averaged delay between 3 and 4 seconds for valence and arousal, and between 5 and 6 seconds for expectancy and power. The differences between those delays could be explained by the higher complexity of human evaluation for expectancy and power.

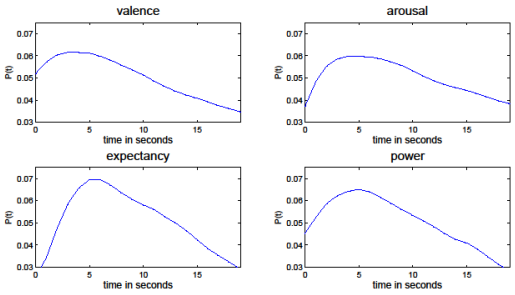

Figure 3: Delay probability distributions

Correlation-based feature selection

We present in this section a feature selection method adapted to a possibly time-delayed label.

The kernel regression proposed in this paper uses a similarity measure based on distances between samples. Using all the features (including the ones that are not useful for our prediction) would corrupt the regression by adding an important noise. We need to identify the most relevant ones and then reduce the number of features that will be used in our distance.

In order to only select the features that are correlated to the label knowing the delay probability distribution, we introduce a time-persistent-correlation-based measure:

$$\rho(f_i(t), y(t)) = \int_{-\infty}^{\infty} r(f_i(t), y(t-\tau))P(\tau)d\tau$$

This way, we consider the correlation between the feature and the label, but also between the feature and different delayed versions of the label weighted by an estimation of the delay probability. As before, with different separate video sequences, we need to calculate the mean of this measure for the different sequences to obtain a correlation score between the i^{th} feature and the label. To simplify notations, we refer to this score as $\rho(f_i(t),y(t))$. This measure is more robust than a simple correlation measure in the case of possibly time-delayed label. By selecting features maximizing $\rho(f_i(t),y(t))$, we select a relevant set of features.

Clustering

We present in this paragraph a clustering step with supervised weighted-distance learning. The feature selection step presented in the previous paragraph gives a correlation score between the label and each selected feature. We use these scores as the weights of a diagonally-weighted distance d_w, defined as follows:

$$d_w(X,Y) = \sqrt{X^\top W Y}$$

with W a matrix which components are defined as:

$$W_{ij} = \rho(f_i(t), y(t))\delta_{ij}$$

We perform a k-means clustering algorithm to reduce the uncertainty of the label by grouping samples that are close in the sense of the learned distance d_w. We calculate the label of each group as the mean of the labels of the group. In order to initialize the algorithm, we sort out the samples by label values and gather them in k groups of the same size. We calculate the initialization seeds as the means of the features of each group's samples. This initialization is done to ease the repeatability of the clustering and because we expect to gather samples with neighboring labels after the clustering algorithm by using the learned distance d_w. This step leads to the identification of a set of representative samples.

Kernel regression

After these learning steps, the prediction is done by a kernel regression using the Nadaraya-Watson estimator [15]. We use a radial basis function (RBF) combined with the previously learned weighted-distance d_w as kernel. Let $\mathbf{x_j}$, $j = \{1, \ldots, m\}$ be the feature vectors of the m representative samples obtained after clustering, and y_j, $j = \{1, \ldots, m\}$, the associated labels. The prediction for a sample s described by feature vector $\mathbf{x_j}$ is given by:

$$\hat{y}(s) = \frac{\sum_{j=1}^{m} K_\sigma(\mathbf{x_s}, \mathbf{x_j})y_j}{\sum_{j=1}^{m} K_\sigma(\mathbf{x_s}, \mathbf{x_j})}$$

where σ is the spread of the radial basis function and K is defined as:

$$K_\sigma(\mathbf{x_s}, \mathbf{x_j}) = e^{-\frac{d_w(\mathbf{x_s}, \mathbf{x_j})^2}{\sigma}}$$

As a final step, we proceed to a temporal smoothing to reduce the noise of the regressor output.

FUSION

Using the regression method described in the previous section, we obtain 16 signals, which are the predictions of the four dimensions using the four different sets of features. In order to fuse these signals and make the final prediction of the four dimensions, we chose to use local linear regressions to estimate linear relationships between the signals and the labels. More precisely, the coefficients of these linear relationships are estimated as the means of the different linear regressions coefficients

weighted by the Pearson's correlation between the predicted signal and the label of each sequence. Let y_i^j, $i=\{1,\ldots,n_s\}$, j={V,A,E,P} be the labels of the n_s video sequences of the learning set. Let S_i , $i=\{1,\ldots,n_s\}$be the matrices containing the 16 predictions of our system on the n_s sequences of the training set (previously standardized). We estimate the four vectors of coefficients α_j of the linear relationships as follows:

$$\alpha_j = \frac{\sum_{i=1}^{n_s} r(\beta_i^j S_i, y_i^j)\beta_i^j}{\sum_{i=1}^{n_s} r(\beta_i^j S_i, y_i^j)}$$

where $\beta_i^j = (S_i^T S_i)^{-1} S_i^T y_i^j$ is the ordinary least squares coefficients vector for sequence i and label j.

We can then calculate our final predictions for the four dimensions p_j , $j =\{V,A,E,P\}$ as:

$p_j = \alpha_j S_t$ where S_t is a matrix containing the 16 standardized predictions of all the regressors on the test sequence we aim to predict.

EXPERIMENTS

In this section, we present some experiments we carried out to evaluate the different key points of our method. In order to be robust in generalization, we chose to optimize the hyperparameters in subject-independent cross-validation (each training partition does not contain the tested subject). We report here the result of the full system (with feature normalization by subject, our time-persistent-correlation measure and our regression framework). The contribution of each key-point is deeply studied in [18]. The next subsection details the results of the challenge data released by the organizers one week before the challenge deadline.

Fusion evaluation

The proposed fusion method, which is based on a simple linear combination of the inputs learned via local linear regressions, is particularly fast and well-suited for a real-time system.

	Val	Aro	Exp	Pow	Mean
S	0.319	0.538	**0.365**	0.429	0.413
GA	0.281	0.498	0.347	0.431	0.389
LA	**0.354**	0.470	0.323	0.432	0.395
A	-0.057	0.445	0.280	0.298	0.241
F	0.350	**0.644**	0.341	**0.511**	**0.461**

Table 1: Pearson's correlations averaged over all sequences of the AVEC'12 development set.

To evaluate the efficiency of this fusion method and the contribution of each feature set, we present the results we obtained by learning on the training set and testing on the development set in Table 1. Results are given in terms of correlation for valence (V), arousal (A), expectancy (E) and power (P. We also indicate the mean correlation of these four dimensions. S corresponds to the shape features. GA to the global appearance features. LA to the local appearance features and A to the audio features. F corresponds to the fusion.

Results on the test set

We learned our system on the concatenation of the training and the development sets to compute our predictions on the test set. We compare in Table 2 our results to those given in the baseline paper [26]. We can notice that the results obtained on the test set are quite similar to those obtained on the development set. This highlights the high generalization power of the proposed framework. It can be explained by the small number of representative samples for the kernel regression (60 in our system) which limits the flexibility of the model and allows the system to only capture important trends in the data.

	Val	Aro	Exp	Pow	Mean
Our method	**0.341**	**0.612**	**0.314**	**0.556**	**0.456**
Baseline	0.146	0.149	0.110	0.138	0.136

Table 2: Pearson's correlations averaged over all sequences of the AVEC'12 test set.

META-ANALYSIS OF THE CHALLENGE

Challenger results in terms of mean correlation and root mean squared error are compared to the baseline in Fig. 4. The proposed system gets the highest accuracy on both measures.

The system described in [28] extracts and merges visual, acoustic and context relevant features. They propose a method that adapts to the morphology of the subject and is based on an invariant representation of facial expressions. It relies on 8 key expressions of emotions of the subject. In their 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 (particularly smile) in a video sequence is performed in this space with a basic intensity-area detector. The features extracted from audio mode come from the analysis of the speaking turns, sentences and keywords. It is possible, with the transcripts of a conversation, to automatically find the emotional agent of the sequence. Knowing that each agent has its own emotional profile and that most of the time, subject and agent are emotionally synchronized, it's easy to deduce a statistical mean value of the subject's valence and arousal for the sequence. To fuse multimodal features, they use a classical Mamdani Fuzzy Inference System. The results show that the duration of high intensity smile is an expression that is meaningful for continuous valence detection. The main variations in power and expectancy are given by context data. The mean correlation is not far from our (0.43 instead of 0.46) and the root mean squared error is lower.

The third challenger proposal use temporal statistics of texture descriptors extracted from facial video, a combination of various acoustic features, and lexical features to create regression based affect estimators for each modality. The single modality regressors are then combined using particle filtering, by treating these independent regression outputs as measurements of the affect states in a Bayesian filtering framework, where previous observations provide prediction about the current state by means of learned affect dynamics. Tested on the Audio-visual Emotion Recognition Challenge dataset, their filtering-based multi-modality fusion achieves correlation performance of 0.344.

Comparing the three first challengers is interesting on two aspects. First, these results are not "so" good as mean correlation is always lower than 0.5. Looking at predictions in detail, we can see that sometime, they are quite perfect and sometime, completely at the opposite of ground truth. Let us notice that the analysis of ground truth labels [28] shows that the mean correlation between annotators is 0.45. Given both results, we may ask if we face an ill posed problem! Maybe affective states are too subtle to be detected by using only the audio-visual channels. Nevertheless, each challenger uses (mostly) different cues and gets more or less accuracy on the four affective dimensions. We can guess that selecting or combining in some way all these cues should lead to better results.

CONCLUSION

We presented a complete framework for continuous prediction of human emotions based on features characterizing head movements, face appearance and voice in a dynamic manner by using log-magnitude Fourier spectra. We introduced a new correlation-based measure for feature selection and evaluated its efficiency and robustness in the case of possibly time-delayed labels. We proposed a fast regression framework based on a supervised clustering followed by a Nadaraya-Watson kernel regression. Our fusion method is based on simple local linear regressions and significantly improves our results. Because of the high power of generalization of our method, we directly learned our fusion parameters using our regressors outputs on the training set. In order to improve the fusion for methods that are more sensitive to over-fitting, we would have to learn these parameters in cross-validation. Some modifications on our system would be needed to increase its performance regarding this measure. The SEMAINE database on which our system has been learned and tested contains videos of natural interactions but recorded in a very constraint environment. A perspective for adapting these kinds of human emotion prediction systems to real conditions, as for Intelligent Tutoring Systems, would be to learn the system on "in the wild" data.

BIBLIOGRAPHY

1. Cruz, B. Bhanu, and S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 341-350, 2011.
2. M. Dahmane and J. Meunier. Continuous emotion recognition using gabor energy _lters. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 351{358, 2011.
3. S. Dawson, L. Heathcote, G. Poole. Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience. International Journal of Educational Management, 24(2): 116 – 128, 2010.
4. P. Ekman. Facial expression and emotion. American Psychologist, 48(4):384, 1993.
5. P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial action. Manual for the Facial Action Coding System, 1978.
6. J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. Psychological science, 18(12):1050, 2007.
7. R. Greenleaf. Motion and Emotion in Student Learning. Principal Leadership, 2003.
8. H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In Proc. of Intelligent Virtual Agents (IVA'10), pages 371{377, 2010.
9. H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In Proc. IEEE Int'l Conf. Face & Gesture Recognition (FG'11), pages 827-834, 2011.
10. B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In Proc. IEEE Int'l Conf. Face & Gesture Recognition (FG'11), pages 314-321, 2011.
11. Kort, B., Reilly, R., & Picard, R. W.. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In Proc. of the IEEE Int'l Conference on Advanced Learning Technologies, pages 43-46, 2001.
12. Computer Society Press, 43-46. D. McDuff, R. El Kaliouby, K. Kassam, and R. Picard. Affect valence inference from facial action unit spectrograms. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops (CVPRW'10), pages 17-24, 2010.
13. G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME'10), pages 1079-1084, 2010.
14. H. Meng and N. Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 378-387, 2011.
15. E. Nadaraya. On estimating regression. Theory of Prob. and Appl., 9:141{142, 1964.
16. M. Nicolaou, H. Gunes, and M. Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, pages 43-48, 2010.
17. M. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. Image and Vision Computing, 30(3), 186-196, 2012.

18. J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, Robust continuous prediction of human emotions using multiscale dynamic cues, In Proc. of International Conference on Multimodal Interaction, pages 501-508, 2012.

19. Picard, R. W.. Affective computing. Cambridge: MIT Press, 1997.

20. Paramythis, A., & Loidl-Reisinger, S.. Adaptive Learning Environments and e-Learning Standards**.** In R. Williams (Ed.), Proceedings of the 2nd European Conference on e-Learning (ECEL2003), pages 369-379, 2003.

21. G. Ramirez, T. Baltrusaitis, and L. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 396-406, 2011.

22. J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. International Journal of Computer Vision, 91(2):200-215, 2010.

23. A. Savran, H. Cao, M. Shah, A. Nenkova, R. Verma, Combining Video, Audio and Lexical Indicators of Affect in Spontaneous Conversation via Particle Filtering , In Proc. of International Conference on Multimodal Interaction, pages 485-492, 2012.

24. A. Sayedelahl, P. Fewzee, M. Kamel, and F. Karray. Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 407-414, 2011.

25. B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011: the first international audio/visual emotion challenge. In Proc. of Affective Computing and Intelligent Interaction (ACII'11), pages 415-424, 2011.

26. B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge – an introduction. In Proc. of International Conference on Multimodal Interaction, pages 361-362, 2012.

27. T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 42(4):993-1005, 2012.

28. C. Soladie, H. Salam, C. Pelachaud, N. Stoiber, R. Seguier, A Multimodal Fuzzy Inference System using a Continuous Facial Expression Representation for Emotion Detection, In Proc. of International Conference on Multimodal Interaction, pages 493-500, 2012.

29. M. Valstar, M. Mehu, B. Jiang and M. Pantic, Meta-Analysis of the First Facial Expression Recognition Challenge, Transactions on Systems, Man, and Cybernetics, Part B, 42(4):966-979, 2012.

30. M. Wollmer, M. Kaiser, F. Eyben, and B. Schuller. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 2012.

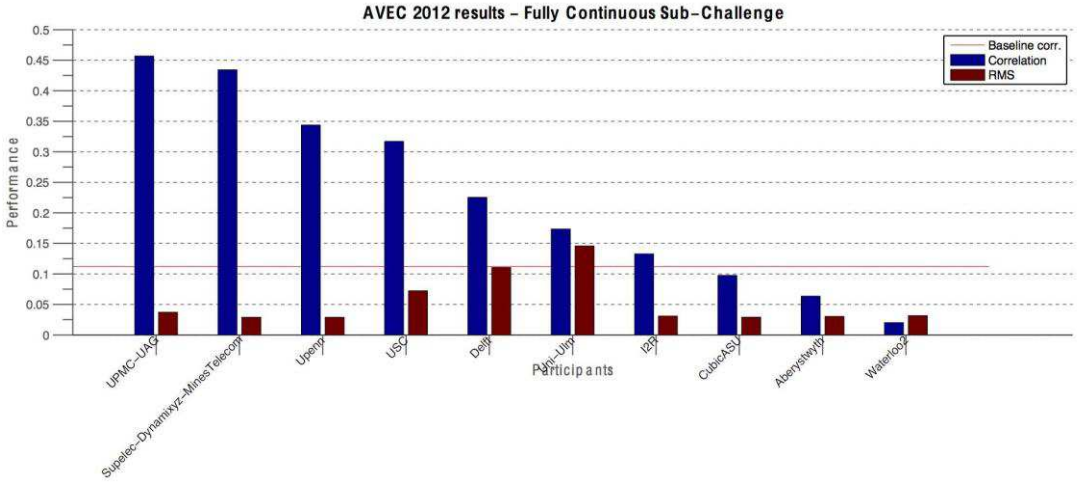

Figure 4: Baseline correlation, mean correlation and root mean squared error of the ten challengers