Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 120 (2015) 556 – 559

Procedia Engineering

www.elsevier.com/locate/procedia

EUROSENSORS 2015

Continuous monitoring of emotions by a multimodal cooperative sensor system

Arianna Mencattini[a], Fabien Ringeval[b], Björn Schuller[b,c], Eugenio Martinelli[a] and Corrado Di Natale[a]

[a]*Dept. Elect. Engineering, University of Rome Tor Vergata, Roma, Italy*
[b]*Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany*
[c]*Department of Computing, Imperial College London, London, U.K*

Abstract

Multimodal emotion recognition is a challenging topic that aims at determining the affective state of a subject by combining audio-visual and physiological signals acquired in a naturalistic environment. This procedure can be used to monitor the emotional state of a subject affected by mental disorder or under medical treatment. Common attempts principally learn a unique complex machine learning system on descriptors collected from different subjects. The novel paradigm of single-subject multimodal regression model (SSMRM) that we propose in this study is embedded in a averaging-based merging strategy that aggregates the responses provided by each model during the test of a new subject. This new approach presents a flexible architecture able to continuously embed new models without global re-training.

Peer-review under responsibility of the organizing committee of EUROSENSORS 2015

Keywords: Speech emotion recognition, multimodal cooperative sensorial systems, naturalistic emotional display

1. Introduction

Emotions have deep impacts in human perception, but also in everyday learning and decision-making strategies. Despite the fact that cognitive processes used to encode affective information during social interactions are relatively complex, humans can easily manage to decode such information in real time from multimodal cues. On the contrary, the effort required to computer-based systems for a reliable and autonomous understanding of emotion is still challenging, above all for multimodal emotion recognition.

Nowadays, clinical applications exploit the benefits of pervasive computational systems to complement standard medical practices, through on-line monitoring of patient's emotions, such as arousal (degree of stimulation) or

Peer-review under responsibility of the organizing committee of EUROSENSORS 2015
doi:10.1016/j.proeng.2015.08.716

valence (degree of pleasantness) [1]. Humans can express their emotional state through the combination of three modalities: acoustic, visual and physiological. Unfortunately each modality has its own intrinsic relationship with the emotional content so that feature selection is crucial to identify the most relevant descriptors of a given emotional display [2,3]. To this regards, in this paper, we propose a new tightly coupled multisensory data fusion, as an extension of our previously developed emotion recognition system in a context-dependent fashion [4].

2. Methods and Results

The system is based on the extraction of a set of 284 multimodal features computed on 10 different subjects from the RECOLA database [5], a new multimodal corpus of spontaneous affective interactions in French. The feature set includes 160 acoustic descriptors [6], 40 facial descriptors [7,8], 54 descriptors of the electrocardiogram (ECG) and 60 descriptors of the electro-dermal activity (EDA) [9,10]. As shown in Fig. 1, each single-subject multimodal regression model (SSMRM) is trained and optimized on the multimodal sensor data continuously annotated by experts in terms of arousal and valence dimension.

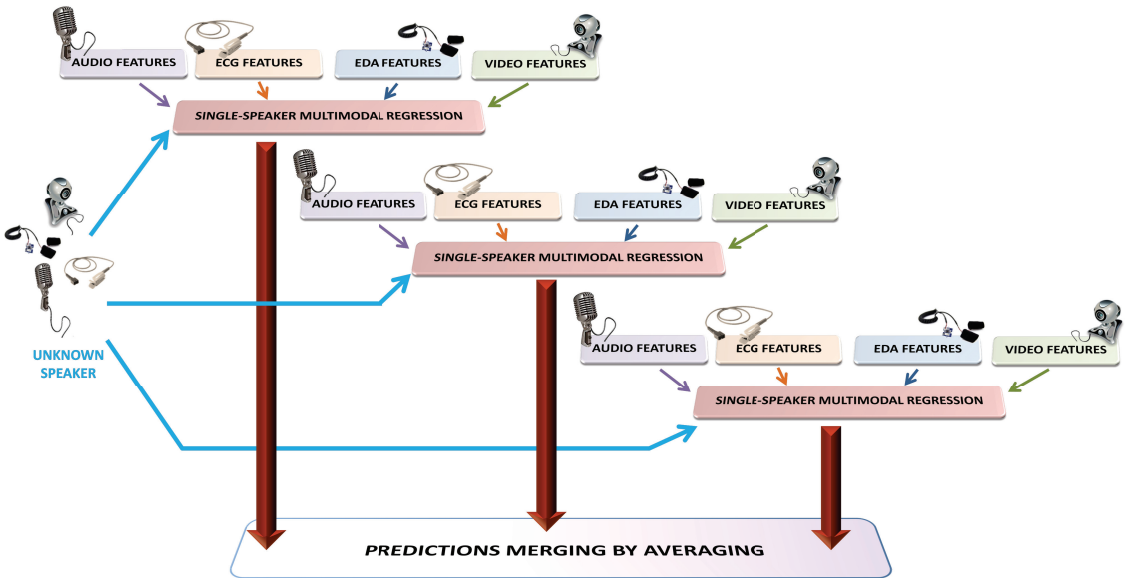

Figure 1: The overall architecture of the proposed multimodal cooperative sensorial system. Modalities can be in turn removed for implementing the seven outlined case studies.

During the test phase (blue arrows), each SSMRM is applied on the multimodal features computed for the new subject and the different responses are averaged to provide a unique estimation of his/her emotional state. Performance evaluation is carried out in a leave-one-speaker-out (LOSO) cross-validation procedure.

Considering the four-dimensional multimodal framework, we compare the results obtained in seven different case studies. The first three cases are used to demonstrate the benefits of a multimodal prediction, whereas the remaining four cases are used to compare the model with a single modal prediction based on each of the four modalities. Performance are evaluated with the Pearson's Correlation Coefficient (CC) between the prediction and the time-continuous rating of arousal or valence provided by six observers, and are reported in Fig. 2.

It is worth to mention that in these applications, a CC around 0.4 for valence may be considered as a positive result. [9]. Noticeable is that the addition of video cue improves the overall results obtained using audio, especially in valence dimension.

Since performance vary over different subjects, occasionally also ECG and EDA cues introduce improvement. Figure 3 reports an example of arousal prediction.

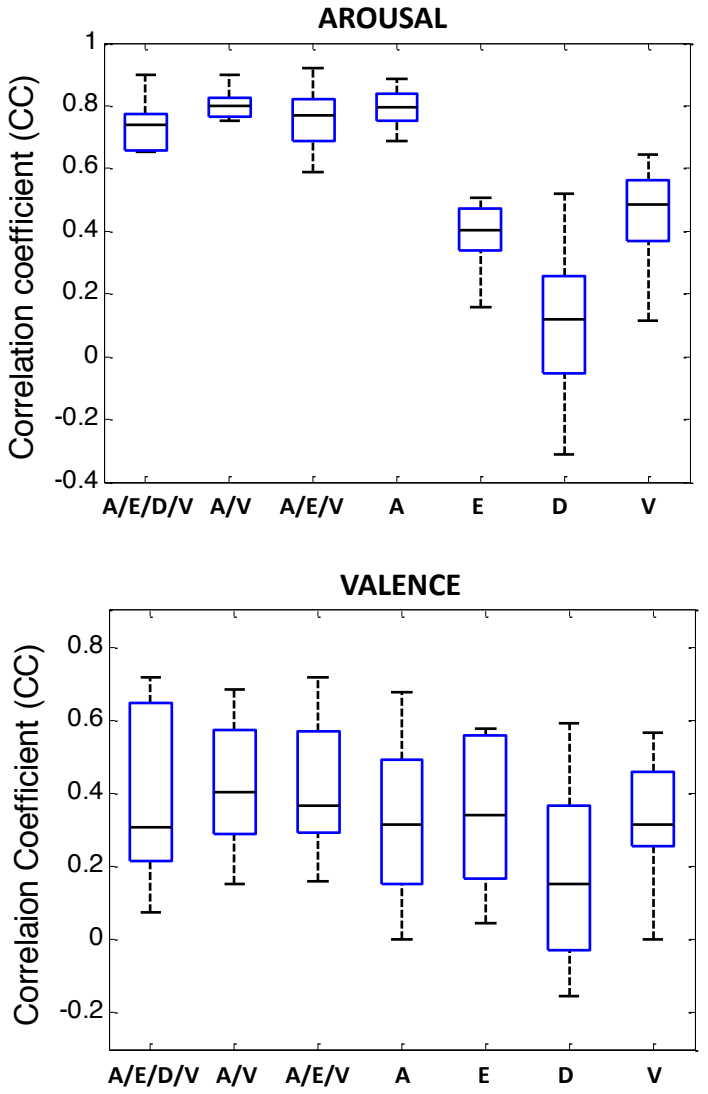

Figure 2: Box-plot of the correlation coefficient between the prediction and the gold standard; arousal (up) and valence (bottom) for multimodal and monomodal configurations, A: audio, E: ECG, D: EDA, V: video, as well as the following combinations: A/E/D/V, A/V, and A/E/V.

Figure 3 reports an example of arousal prediction. The multimodal scenario presents an accurate linear and time continuous emotion prediction using the fusion of the three modalities (audio, ECG, and video) (A/E/V). The modularity of the system and the merging strategy of multiple individually trained predictions is able to ensure a high flexibility of the architecture: as an example additional single models may be added without global re-training making it feasible to continuously monitor the affective state of a patient, as adjuvant practice to therapies.

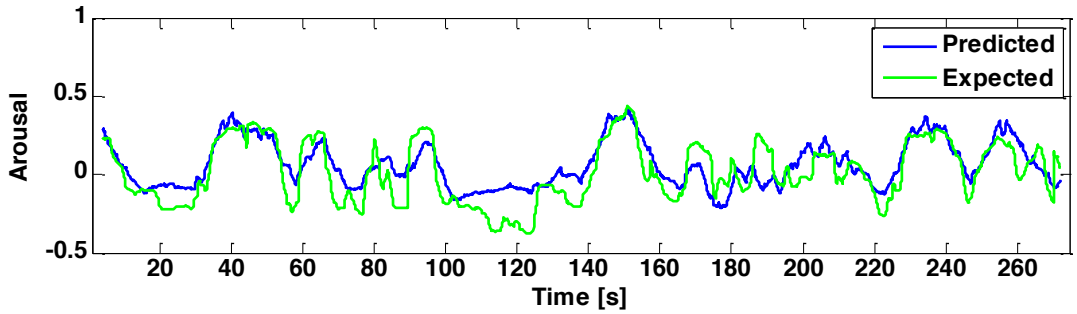

Figure 3: Example of predicted arousal using the best combination of modalities: audio, electrocardiogram and video (A/E/V).

3. Conclusion

In this work, continuous monitoring of emotions by a multimodal cooperative sensor system is presented and applied to the naturalistic emotional database RECOLA. Ten independently trained SSMRMs were learnt from 284 heterogeneous descriptors extracted from four different modalities (AUDIO, EDA, ECG; and VIDEO). Then, during the test of a new speaker, the predictions provided by applying each single SSMRM are averaged to provide a unique estimation of the unknown emotional state of the subject. Combinations of the four different modalities have been compared with the four single modality systems to demonstrate the effectiveness of the multimodal approach. Results show that the technique is promising for diversified scenarios: first, he proposed architecture is perfectly suitable for mobile applications, thanks to the easiness and the flexibility to develop single models separately trained on distinct sequences with different emotional contents. Also, web-based applications could offer the possibility to everyone to upload to the cloud his/her speech sequence along with the corresponding annotation. Remote care assistance and patient-centered applications may benefit of such kind of auxiliary examination of the disease.

4. References

[1] N. Sebe, I. Cohen T. S. Huang (2005): Multimodal emotion recognition. Handbook of Pattern Recognition and Computer Vision 4, 387-419.

[2] J. Wagner, E. André, F. Jung (2009): Smart sensor integration: A framework for multimodal emotion recognition in real-time. Affective Computing and Intelligent Interaction and Workshops, ACII 2009. 3rd International Conference on. IEEE.

[3] J. Kortelainen, S. Tiinanen, X. Huang, X. Li, S. Laukka, M. Pietikainen, T. Seppanen (2012): Multimodal emotion recognition by combining physiological signals and facial expressions: a preliminary study. In Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE (pp. 5238-5241).

[4] A. Mencattini, M. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, C. Di Natale (2014), Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. Knowledge-Based Systems, 63, 68-81.

[5] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, (2013): Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, Proc. of Face & Gestures 2013, 2nd IEEE Inter. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), Shanghai, China.

[6] F. Eyben, F. Weninger, F. Gross, and B. Schuller (2013): Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM Multimedia (MM)*, Barcelona, Spain, 2013, pp. 835–838.

[7] X. Xiong and F. De la Torre (2013): Supervised descent method and its applications to face alignment. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE.

[8] A. Yuce, M. Sorci, J.-P. Thiran (2013): Improved local binary pattern based action unit detection using morphological and bilateral filters." Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE.

[9] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller (2014). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data, Pattern Recognition Letters. Article in Press.

[10] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras (2012), DEAP: A database for emotion analysis; Using physiological signals, IEEE Transactions on Affective Computing, 3 (1), pp. 18-31.