

Title: Multimodal emotion recognition: Implementation, results and challenges

Authors: ERMIS partners (KCL, QUB, ICCS)

This poster builds on the concepts, results and experiences obtained during the FP5 IST ERMIS project on emotional state recognition, presenting a number of limitations that have been identified and challenges that have arisen. ERMIS adopted both a unimodal approach on emotion recognition, based on facial expression analysis, as well as a multimodal approach, based on both facial and vocal cue analysis.

In the unimodal case, a rule-based system for emotion recognition was created, based on MPEG-4 concepts, classifying the user's emotional state in terms of six universal, or archetypal, expressions (joy, surprise, fear, anger, disgust, sadness). The initial rules were extended to deal with the continuous emotional analysis task using a variety of clustering and fuzzy reasoning techniques. Results indicated the good performance and potential of the developed technologies, especially in the case of extreme and/or well-defined facial expressions.

In multimodal emotion recognition an attention-based neural network was generated, using features from various modalities that correlated with the user's emotional state. These features were fed to a hidden layer representing the emotional content of the input message. Attention acted as a feedback modulation onto the feature inputs, so as to amplify or inhibit the various feature inputs, as they are or are not useful for the emotional state detection. The basic architecture was thus based on a feed-forward neural network with the addition of a feedback layer modulating the activity in the inputs to the hidden layer. Results have produced multi-modal time series streams of text, prosodic features and facial features.

ERMIS was a pioneer project in extensive multimodal analysis of spontaneous emotional discourse. The project managed to provide a coherent environment for emotion recognition based on speech and facial input analysis.

A challenge that has arisen from this work is the investigation of the relation between vocal and facial cues in multimodal analysis. Up till now only parts of speech, comprising meaningful utterances, were coupled with a choice of facial expressions; the exact coupling of phonemes with the corresponding visemes (facial equivalents of phonemes) is a challenge that awaits to be met. Adopting such an approach will be the first step in dealing with the recognition process through time and not only in isolated time intervals or image snapshots, such as those used in

visual recognition. Further more, the inclusion of extra information such as gesture analysis, could shed more light in the emotional state recognition process.

Another challenge identified is the coupling of emotional states observed, with stimuli that elicit them. An episode-driven study of emotional expressions could help clarify emotional behaviour in relation to cognition and action. The exploration of interactive scenarios both in human-human and human-computer situations can explain the mechanisms of emotional behaviour in conjunction with motivation and goal attainment. Such research can help in bridging the existing problematic gap between the signal processing level of emotion recognition and the psychological/ cognitive approach.