Implications of Multimodal Sensing for an Emotion Markup Language

Christian Peter[1, 2]

[1] Fraunhofer IGD, Joachim-Jungius-Str. 11, 18059 Rostock, Germany
christian.peter@igd-r.fraunhofer.de
[2] Graz University of Technology, Inffeldgasse 16c, A-8010 Graz, Austria
c.peter@cgv.tugraz.at

Abstract. This paper wants to raise awareness of the specific issues posed by multimodal sensing of affective states. It first sketches typical scenario for multimodal sensing of affective states and draws conclusions for an emotion markup language. The issues identified in this paper have their origin mainly in the often very different nature of modalities and that technological developments introduce new possibilities to observe a person's affective state. It is hence considered necessary for an emotion markup language to provide for a high degree of freedom for manufacturers of affect sensors and developers of affective systems while also guaranteeing unambiguous annotation of affective states.

Keywords: Emotion, Annotation, Multimodality, Affective Sensors.

1 Introduction

This paper sketches a simple, yet typical scenario for multimodal sensing of affective states and draws conclusions for an emotion markup language. More complex scenarios than that given here can be envisioned and will in fact become reality in the near future: smart homes with ambient sensors and ambient interaction capabilities are being developed in several current European projects; first solutions of smart living rooms are already deployed in residential homes; smarter work environments for e.g. knowledge workers will also include means to infer the employee's current mental state to provide just the information needed in the most appropriate way. Multimodal sensing is central to all these endeavours.

Multimodality in this context means the use of multiple modalities (or information "channels") to get information on the current affective state of a person. While usually one sensor is considered to provide for one modality (e.g. a biosensor describing the physiological modality), it must be considered that one device might serve multiple modalities, just as a camera can be input device for both, gestural and postural information. Also it should be mentioned that sensor fusion techniques allow creating fused modalities of higher robustness. For instance, a fused modality "activity" could be fed by a camera plus an acceleration sensor worn by the person.

Further it should be mentioned that technological developments in the sensor domain allow for accessing more and more emotion-related signals (e.g. near-infrared

imaging techniques) and that research in human behaviour might lead to new modalities to look at. For instance, while gesture, gait and posture are the most commonly observed modalities related to behaviour, accelerometer-based activity information might open a new information channel for also inferring on a person's affective state.

The following section shortly describes a use case in which multimodal sensing for affective states is used. The next section then highlights implications of multimodal sensing for an emotion markup language. A short summary is given at the end.

Use case: User Experience and Usability Evaluation

Karina works as a user experience consultant and evaluates the usability of diverse software products. In her lab, she uses a system that tracks a user's behaviour while working with software programs or using web pages.

Besides capturing the content of the computer screen and logging keyboard and mouse events, this new system also collects emotion information on the user while interacting with the product of interest, by use of several sensing technologies. The system is equipped with various sensors for both, behaviour tracking and emotion detection, like the following:

- Eye gaze tracker for tracking the visual focus of the user;
- Camera for observing facial features, head gestures and body posture;
- Microphone for capturing speech (for her main customer, a telecoms company);
- Sensors for physiological parameters (skin conductivity, skin temperature, pulse, respiration).

In a typical evaluation session, a test user performs a certain task on the software to be evaluated. The system's unobtrusive sensors monitor the behaviour and observe emotion-related signs:

- The mouse movements, clicks, and keyboard strikes;
- The trace of the visual focus;
- Changes in facial expressions, head movements like nods or head shakes, and body movements like leaning for- or backwards;
- The tone of the voice;
- Changes in emotion-related physiological parameters.

During the session, in the adjacent room Karina observes the scenery at a second screen and adds annotations to the data stream. Those annotations are e.g. on the user's performance, Karina's impressions of the user's current feelings and likely user experience, and environmental events like distracting sounds or other observations she makes. After the test, Karina talks with the person about experiences made, likes and dislikes on the software, and feelings at particular situations. For this, she uses the playback feature of the analysis tool and the markers set by her manually or by the system based on on-the-fly data analysis and classification results.

All sensor information collected are automatically analysed and classified by the software. The program analyses each "channel" not just for interactions, but also for

signs of affective states and changes thereof. The channels (face camera, speech data, physiology, body posture) are analysed individually for their affective content. In a next step, these classification results are merged and a summarized affective state estimate is calculated. Important for Karina is that both levels are available for her to work with: the channel-specific emotion classification as well as the machine's summary. Particularly the channel-specific markers set by the system give her interesting hints on emotional episodes during the usability test. This allows her to better understand why users act the way they do at certain situations while also comparing her own judgement to that of a second, "neutral observer".

Implications of Multimodal Sensing of Affective States

The example given above is a typical scenario for multimodal sensing of affective states. Different sensors observe a person performing a task and draw their own conclusions on her current affective state. Moreover, one device, the camera, serves as input for multiple channels: facial features, head gestures, and posture. All modalities are read out permanently while certainly not all will provide affective information all the time (physiology being an exception). For instance, facial features will quite likely be "neutral" most of the time as people tend not to display much emotion through the face when interacting with machines. Voice data might not be present at all in some scenarios, while in others they might be the main information source, for instance when testing help desk software. Combining the classification results of the individual channels hence seems to be requisite for having reliable emotion information available most of the time.

Issues arising

The example above addresses several issues relevant for an emotion markup language.

Indicating the modality through which an emotion has been observed: As has been illustrated, different channels, or modalities, might provide different information on observed emotions. It is hence mandatory in multimodal settings to always provide information about the modality that provided the information. Please note again that "modality" is not necessarily the same as "sensor".

Allow for occasionally present modalities: In real life, modalities that provide emotion information permanently are an exception. Apart from physiological sensors, most modalities will not permanently deliver usable information either because of adverse observation conditions (e.g. person out of sight, bad lighting), or simply because there is no observable sign for them (e.g. a neutral face, no speech). As a consequence, it should be possible to add emotion information from modalities as they occur. It should also be easy to add modalities to a given set.

Allow for multiple emotion annotations for a given time interval: Since different modalities might lead to different conclusions on a person's emotional state, it must be possible to provide different emotion information for a given episode.

Time information should be unambiguous: Apart from other obvious reasons for unambiguous time information, multimodality brings with it that different modalities provide their information with certain (different) delays. That means that for one and the same observed episode emotion information will be provided several times, with delay, from several modalities. The markup language should allow for this and provide means to unambiguously match those observations.

Provide confidence information: Emotion estimates of different modalities differ in their accuracy depending on environmental factors, behaviour of the person, and technical issues. For instance, when using facial features as well as speech analysis, the facial emotion estimate is less reliable than the speech analysis results for episodes in which the person talks. To take this fact into consideration when calculating a summarized emotional state, reliability information should be provided for each emotion estimate of each modality. In the given example, the facial feature classifier would still provide an emotion estimate, accompanied by a fairly low confidence value.

Be independent from sensor-specifics in the emotion description: Sensor specifics (video frame rates, sample rates, thresholds, data enhancement information) should be hidden in the emotion description part. The information on the emotion should be uniform for all modalities.

Allow for sensor-specific information (elsewhere): Information on the sensor, data enhancements performed and classification methods used might be crucial for consecutive processing steps. It must hence be possible to provide such information for each modality used.

Allow for different emotion models between modalities: it is quite likely that different sensing components (made by different manufacturers) will use different emotion models (2 or 3 dimensions, Ekman categories, Frijda's categories, own ones). This must be possible.

Allow for user-defined mapping of emotion models: Since different modalities might use different emotion models to communicate their emotion information, it must be possible to define application-specific mappings of those.

These are major issues raised by multimodal sensing of affective states without being considered an exhaustive list. With the technological development in the sensor and data analysis domain developing fast, it is important to not limit the number or nature of modalities accepted by any emotion annotation language. The language should give users and system designers most possible freedom to set up their affective system as they desire.

Summary

This paper sketched a typical scenario of multimodal affect sensing. Multimodality poses several challenges for annotating, storing, and communicating affective states. The reason lies mainly in the fact that more and more modalities are utilized to infer affective states of persons and that no general "best" way of representing emotions has been found so far. It is hence necessary to provide for a high degree of freedom for manufacturers of affect sensors and developers of affective systems.