Social Signal Interpretation (SSI)

A Framework for Real-time Sensing of Affective and Social Signals

Johannes Wagner · Florian Lingenfelser · Nikolaus Bee · Elisabeth André

Abstract The development of anticipatory user interfaces is a key issue in human-centred computing. Building systems that allow humans to communicate with a machine in the same natural and intuitive way as they would with each other requires detection and interpretation of the user's affective and social signals. These are expressed in various and often complementary ways, including gestures, speech, mimics etc. Implementing fast and robust recognition engines is not only a necessary, but also challenging task. In this article, we introduce our Social Signal Interpretation (SSI) tool, a framework dedicated to support the development of such online recognition systems. The paper at hand discusses the processing of four modalities, namely audio, video, gesture and biosignals, with focus on affect recognition, and explains various approaches to fuse the extracted information to a final decision.

Keywords Social signal processing · Human-centred computing · Affective computing · Multimodal fusion · Machine learning · Real-time recognition

1 Introduction

Today's human computer interaction (HCI) still requires the user to adapt to the computer, e.g. using a keyboard to type commands or using a mouse to steer an application. Often, this hampers the interaction with a system, especially for non-skilled users. Moving towards a more intuitive interaction is therefore an important aim of next-generation HCI. Intuitive interaction requires the computer to correctly interpret and understand signs of human behaviour, which are the foundation of human communication. In other words, an intelligent system should adapt to the communication style of the user, rather than the other way round [23].

During interaction with another human being we use our visual and audible senses to mutually express our needs and goals. In order to achieve a more human-like communication, we need to equip machines with the ability to recognize and interpret the different types of human generated signals, including language, gestures, mimics, emotions, etc. This is a challenging task as human expressions do not follow the precise mechanism of a machine, but are tainted with a high amount of variability, uncertainty and ambiguity. In addition, for most applications a prompt reaction to a user's behaviour is required, which means that processing and interpretation of the captured signals must be done on-the-fly.

Automatic sensing of affective and social signals in real-time systems can generally be reduced to three main tasks: (a) *Data segmentation* is the process of detecting on- and offset of actions, which carry relevant information about the user's intention and goals. (b) *Feature extraction* relates to the reduction of a raw sensor stream to a set of compact features—keeping only the essential information necessary to classify the observed behaviour. (c) *Classification* describes the mapping of observed feature vectors onto a set of discrete states or continuous values. This may also include a so called garbage model, which collects observations that do not fit to the current classification model. The *collection of representative samples* to train the classification model is in fact an important pre-step. Usually, this requires separate

J. Wagner F. Lingenfelser · N. Bee · E. André
Lab for Human Centered Multimedia, Augsburg University, Universitätsstraße 6a, 86159 Augsburg, Germany
e-mail: johannes.wagner@hcm-lab.de

F. Lingenfelser
e-mail: florian.lingenfelser@hcm-lab.de

N. Bee
e-mail: nikolaus.bee@hcm-lab.de

E. André
e-mail: elisabeth.andre@hcm-lab.de

recording sessions during which users are either asked to show certain actions or interact with a system that has been manipulated to induce the desired behaviour. Afterwards, the collected data is observed by annotators, who label the observed user actions.

In the following, we present our Social Signal Interpretation (SSI) tool, a framework dedicated to support the development of online recognition systems [28]. SSI covers the tasks necessary to assemble a complete machine learning pipeline, ranging from live sensor input and real-time signal processing, to model training and online classification. Even though SSI is not limited to the task of emotion recognition it has been developed with a focus on sensing of affective and social signals. In the following sections we illustrate the functionality of SSI by discussing the extraction of affective cues from four modalities, namely speech, face, gestures and biosignals. Afterwards, we explain the different approaches SSI offers to fuse the extracted information to a final decision.

2 Modelling Affect

In order to build a machine that is able to recognize the affect state of a user we need to define an emotional model that can be computed by a machine. Hence, we will shortly introduce the categorical and the dimensional model as two prominent ways to conceptualize the phenomenon of emotion.

Within the categorical model, emotions are described as discrete categories, such as happiness, sadness, fear or anger. Researchers like Ekman have tried to define sets of basic emotions, which are universally valid among all humans [8]. An advantage of a category representation is that the terms are adopted from daily life and hence a common understanding about their meaning exists. An alternate representation of emotion are dimensional models, which define emotions in terms of dimensions. Mehrabian [22], for instance, suggests to characterize emotions along three axes, namely pleasure, arousal and dominance. While this representation is less intuitive, it allows continuous blending between affective states.

Both models are simplified representations of emotions that do not cover all their aspects. Nevertheless, they are useful tools to model emotions in the computational world of a machine. If we group samples by assigning them to discrete categories, we can apply supervised learning techniques to categorize unseen data. A trained classifier associates an unknown sample with the category it fits best. Dimensional models, on the other hand, suit the fusion of competitive decisions from different components well and hence can be used to combine the cues from different modalities to a global decision.

In addition to categorical and dimensional models, appraisal models have been employed to model affect in computational systems. Prominent examples include EMA [12] and ALMA [9]. However, in most cases, they have been used to simulate how an agent appraises situations and events and apart from a few exceptions [7] not been used for recognition tasks. In SSI categorical and dimensional models are used.

3 Affective Recognition

In the following we will describe some of the feature extraction and fusion algorithms that have been incorporated in the SSI framework to derive the affective state of a user. Note that described methods can be plugged to a recognition pipeline that processes live input from a single or multiple sensor devices. SSI automatically handles synchronization between the components of a pipeline and allows several pipelines to run in parallel and share data streams [28].

3.1 Speech

Since we are interested in the affectiveness of the user's voice, we only analyse parts of the audio signals where speech is present. For every speech segment we compute information on the prosody of the voice. Since all features need to be extracted fully automatically and in near real-time, we calculate only acoustic features related to the paralinguistic message of speech (see [26] for more details regarding the extraction of acoustic features). The resulting feature vector is passed to a classifier which assigns a discrete class label to it.

In a previous study the feature set was evaluated on the Berlin Database of Emotional Speech [3] that is commonly used in off-line research (7 emotion classes, 10 professional actors) and achieved an average recognition accuracy of 80%. On the FAU Aibo Emotion Corpus as part of the INTERSPEECH Emotion Challenge 2009 [24], we were able to slightly exceed the baseline given by the organizers for a 5 class problem (anger, emphatic, neutral, positive and rest) [25]. Both corpora have been intensively studied in the past by many researchers working on emotion recognition from speech and serve as a kind of benchmark in this area. In both studies we adopted Naive Bayes as classification scheme.

3.2 Face

To analyse the facial expressions of a user we use the SHORE library [19] (provided by Fraunhofer[1]), which has

[1] http://www.iis.fraunhofer.de/en/bf/bv/ks/gpe/demo/.

been integrated into the SSI framework. For each image frame, SHORE reports the bounding of found faces. It additionally extracts a set of features related to a person's mimic—including position of eyes, nose and mouth. SHORE already reports scores for different facial expressions, namely happy, angry, sad and surprised.

In contrast to the processing of the audio stream, where a classification result is received only after a complete utterance was detected, the video is processed per-frame, i.e. a statement is received for each frame in which a face was detected. In order to smooth the scores and reduce the effect of outliers, we average score values over several frames. Performance of facial features has been tested on the CALLAS Expressivity Corpus [5]. The corpus includes 2251 video samples of 21 users and was recently labelled in terms of arousal and valence dimensions. By applying a realistic leave-one-user-out cross validation we observed 50% class-wise recognition rate for the four quadrants. Again a Naive Bayes classifier was applied.

3.3 Gestures

To capture a user's gestural behaviour, we can rely on 3-axes acceleration sensors by the Wiimote (Wii™). Like with speech, where we distinguish between *what* is said and *how* it is said, gesture analysis can either aim at recognising certain pre-defined gesture commands, or extract movement properties such as relative amplitude, speed, and movement fluidity. While SSI includes an implementation of the so called $1 recognizer to recognize discrete gestures [30], we now focus solely on the analysis of gestural expressivity.

Expressivity parameters were originally defined by Hartman et al. for expressive gesture synthesis for embodied Conversational Agents [13]. Based on their concepts Caridakis and colleagues [4] propose a set of six expressivity parameters, namely overall activation, spatial extent, temporal, fluidity, power/energy and repetitivity, which they apply to hand tracking from video images in order to measure expressivity of gestures. Each parameter captures certain properties that include information on *how* a gesture is performed, e.g. overall activation is considered as the overall quantity of movement, while fluidity differentiates smooth/graceful from sudden/jerky gestures.

To extract same features from the acceleration signals, we first eliminate the influence of gravity by removing the linear trend from each of the three acceleration axis. After that, first derivative is calculated and cumulative trapezoidal numerical integration is applied to deduce velocity and position. Finally we calculate the power from each signal, as well as fluidity from position.

3.4 Biosignals

As a further carrier of emotional information, we consider physiological data, including electrocardiography, electromyography, skin conductance, and respiration. Since physiological reactions of a person are controlled by the autonomous nervous system the physiological reactions of the body are less subdued by the human will and to social masking, moreover they are permanently present and can be captured continuously.

In [16], we proposed a wide range of physiological features from various analysis domains, including time/frequency, entropy, geometry, subband spectra, etc. To learn more about the mapping between the observed patterns and certain emotional states, we conducted several experiments during which we captured the biosignals from users while they were put into different affective states. In one, where music was used to elicit the desired emotions, we were able to distinguish four emotional states (joy, anger, sadness, and pleasure) with an accuracy of over 90% [27].

For online analysis we must rely on features that can be calculated on-the-fly. In order to prove the applicability in on-line applications we developed a generic set of recursively calculated real-time features and tested them on the same database yielding similar results [14]. Based on the best features we developed an online in-vehicle emotive monitoring system within the EU project METABO [18].

3.5 Multimodal Fusion

So far signals have been processed independently of each other. Now, we need a way to combine information extracted from several sources to a single decision. In the optimal case, the modalities will carry complementary information and fusing them improves the robustness of the recognition.

A straightforward way to fuse observed channels is to merge every calculated feature into a single and high dimensional feature set, used for one classifier (feature level fusion). Accumulated features contain a bigger amount of information than a single modality, so increased classification accuracy can be expected. Alternatively, multimodal features can be sub-sampled and used for the training of smaller classification models (e.g. one classifier per modality). Outputs of these models are combined by means of decision level fusion. Various combination rules are available for this task, including weighted and unweighted voting schemes, algebraic combiners of continuous classifier outputs, specialist selection algorithms, constructed lookup tables, etc. Examples of elaborate decision fusion schemes and comparison to simpler combination rules can be found in [20]. It is also possible to tailor the fusion mechanisms to concrete fields of application, e.g. emotion recognition, as can be seen in [17]. Another way of combining outputs of several classification models is meta level fusion: Instead of following predefined rule-sets for fusion, model results are declared as meta data and used for the training of one or more meta classifiers.

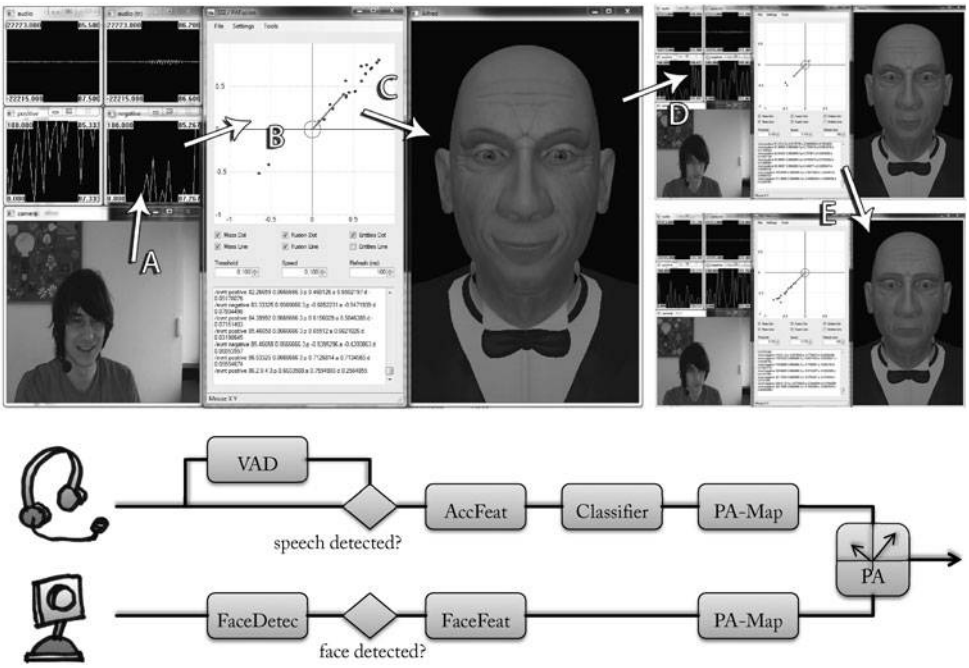

Fig. 1 Affective Listener Alfred: the current user state is perceived using the SSI framework (A); observed cues are mapped onto the PA space (B); PA-values are combined to a final decision and transformed to a set of FACS parameters, which are visualized by Alfred (C); a change in the mood of the user creates a new set of PA values, which moves the fused vector to a new position on the PA space (D); this finally leads to a change in the expression of Alfred (E). A Sketch of the processing pipeline, which analysis the user's speech and mimics, is also shown

A wide range of algorithms for merging multimodal information on described levels are available in the SSI framework. An evaluation and comparison of these methods in terms of recognition performance in relation to single-channel classification can be found in [17]. On the tested corpus we observed an improvement of up to 18% compared to the results from uni-modal classification. In [21] we applied the same fusion methods to a single-channel problem by building tailored feature sets for each target class generated by feature selection. In this way an upgrade on under-represented classes was achieved, which led to an increase in classification performance by up to 5%. Currently, we investigate possibilities for handling temporarily missing modalities in real-time applications, which requires appropriate adaptations of the fusion algorithms.

4 Example Application

SSI has been used in international and national projects as a core component of various affective installations (see e.g. [6, 11, 15, 18]). Details on the integration of SSI within these projects can be found here [29].

In the following, we will briefly describe an exemplary application that was developed with SSI: the Affective Listener "Alfred". Alfred is a butler-like virtual character [1] that is aware of the user and reacts to his or her affective expressions (see Fig. 1). The user interacts with Alfred via spoken natural language. It is important to note that Alfred does not conduct a semantic analysis of the user's utterances, but just tries to determine his or her emotional state from the acoustics of speech and facial expressions. As a response, Alfred simply mirrors the user's emotional state by appropriate facial expressions[2]. This behaviour can be interpreted as a simple form of showing empathy.

In order to recognise the users emotional state, we adopt a modified fusion mechanism described and evaluated in [10]. It is based on a dimensional emotion model, as mentioned in Sect. 2. For the moment, we simplified the model to the pleasure (P) and arousal (A) axis and merge input from only two channels, namely audio and video. If we consider the audio pipeline described in Sect. 3.1 we get as output discrete categories like 'positive-low' or 'negative-high'. Such labels can be directly mapped onto appropriate PA vectors in the according pleasure-arousal quadrant. In case of the video processing described in Sect. 3.2 we get score values that express the confidence on the presence of certain emotional states such as happy or anger. If one of the scores exceeds a certain threshold, we can generate a PA vector of according strength that aims into the quadrant to which the according emotion belongs. Further details on the fusion algorithm can be found in [11]. Note that Gilroy et al. generate one vector per modality, we generate one vector for each detected event. In this way we prevent sudden leaps in case of a false detection in one of the channels. Since the strength of a vector decreases with time, the influence of older events is

[2] In order to illustrate the use of SSI which focuses on emotion recognition tasks, we present a basic version of Alfred here which simply mirrors the user's emotion. Thus, we do not integrate an appraisal model to simulate how Alfred appraises the user's emotional display, but see [2] for a version of Alfred based on the Alma model which combines an appraisal mechanism with a dimensional representation of emotions.

lessened until the value falls under a certain threshold and is completely removed.

5 Conclusion

We have introduced our Social Signal Interpretation (SSI) framework, a tool for the rapid development of online recognition systems to detect social and affective cues from a user. SSI supports input from multiple sensor devices and allows developer to implement processing pipelines by simply plugging together available components. The paper at hand discusses the processing of different modalities (audio, video, gesture and biosignals) with focus on affect recognition. Different ways of fusing information from multiple sensor streams are described, as well. The presented processing and fusion methods are part of SSI and most of the source code is freely available under LGPL.[3]

Acknowledgements The work described in this paper is funded by the EU under research grant CALLAS (IST-34800), CEEDS (FP7-ICT-2009-5) and the IRIS Network of Excellence (Reference: 231824).

References

1. Bee N, Falk B, André E (2009) Simplified facial animation control utilizing novel input devices: a comparative study. In: International conference on intelligent user interfaces (IUI'09), pp 197–206
2. Bee N, André E, Vogt T, Gebhard P (2010) The use of affective and attentive cues in an empathic computer-based companion. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. Benjamins, Amsterdam, pp 131–142
3. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of german emotional speech. In: Proceedings of Interspeech, Lisbon, pp 1517–1520
4. Caridakis G, Raouzaiou A, Karpouzis K, Kollias S (2006) Synthesizing gesture expressivity based on real sequences. In: Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC 2006 conference, Genoa, Italy, 24–26 May
5. Caridakis G, Wagner J, Raouzaiou A, Curto Z, André E, Karpouzis K (2010) A multimodal corpus for gesture expressivity analysis. In: Multimodal corpora: advances in capturing, coding and analyzing multimodality, LREC, Malta, 17–23 May 2010
6. Charles F, Pizzi D, Cavazza M, Vogt T, André E (2009) Emotional input for character-based interactive storytelling. In: The 8th international conference on autonomous agents and multiagent systems (AAMAS) Budapest, Hungary
7. Conati C, Chabbal R, Maclaren H (2003) A study on using biometric sensors for detecting user emotions in educational games. In: Proceedings of the workshop "Assessing and adapting to user attitude and affects: why, when and how?" In conjunction with UM'03, 9th international conference on user modeling
8. Ekman P (1992) An argument for basic emotions. Cogn Emot 6(3):169–200
9. Gebhard P (2005) ALMA: a layered model of affect. In: AAMAS '05: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems. ACM, New York, pp 29–36
10. Gilroy S, Cavazza M, Vervondel V (2011) Evaluating multimodal affective fusion with physiological signals. In: Proceedings of the international conference on intelligent user interfaces. Stanford University, Palo Alto
11. Gilroy SW, Cavazza M, Niiranen M, André E, Vogt T, Urbain J, Seichter H, Benayoun M, Billinghurst M (2009) Pad-based multimodal affective fusion. In: Affective computing and intelligent interaction (ACII), Amsterdam
12. Gratch J, Marsella S (2004) A domain-independent framework for modeling emotion. Cogn Syst Res 5(4):296–306
13. Hartmann B, Mancini M, Pelachaud C (2006) Implementing expressive gesture synthesis for embodied conversational agents, vol 3881, pp 188–199
14. Hönig F, Wagner J, Batliner A, Nöth E (2009) Classification of user states with physiological signals: on-line generic features vs. specialized. In: Stewart B, Weiss S (eds) Proceedings of the 17th European signal processing conference (EUSIPCO), Glasgow, Scotland, pp 2357–2361
15. Jacucci G, Spagnolli A, Chalambalakis A, Morrison A, Liikkanen L, Roveda S, Bertoncini M (2009) Bodily explorations in space: social experience of a multimodal art installation. In: Proceedings of the 12th IFIP TC 13 international conference on human-computer interaction: part II, INTERACT '09. Springer, Berlin, pp 62–75
16. Kim J, André E (2008) Emotion recognition based on physiological changes in music listening. IEEE Trans Pattern Anal Mach Intell 30:2067–2083
17. Kim J, Lingenfelser F (2010) Ensemble approaches to parametric decision fusion for bimodal emotion recognition. In: Int conf on bio-inspired systems and signal processing (Biosignals 2010)
18. Kim J, Ragnoni A, Biancat J (2010) In-vehicle monitoring of affective symptoms for diabetic drivers. In: Fred JFA, Gamboa H (eds) Int conf on health informatics (HEALTHINF 2010), BIOSTEC. INSTICC Press, Valencia, pp 367–372
19. Küblbeck C, Ernst A (2006) Face detection and tracking in video sequences using the modifiedcensus transformation. Image Vis Comput 24:564–572
20. Lingenfelser F, Wagner J, TVJK, André E (2010) Age and gender classification from speech using decision level fusion and ensemble based techniques. In: INTERSPEECH 2010
21. Lingenfelser F, Wagner J, Vogt T, Kim J, André E (2010) Age and gender classification from speech using decision level fusion and ensemble based techniques. In: INTERSPEECH 2010
22. Mehrabian A (1995) Framework for a comprehensive description and measurement of emotional states. Genet Soc Gen Psychol Monogr 121(3):339–361
23. Pantic M, Nijholt A, Pentland A, Huang TS (2008) Human-centred intelligent human-computer interaction (hci): how far are we from attaining it? Int J Auton Adapt Commun Syst 1(2):168–187
24. Schuller B, Steidl S, Batliner A (2009) The INTERSPEECH 2009 emotion challenge. In: ISCA (ed) Proceedings of Interspeech 2009, pp 312–315
25. Vogt T, André E (2009) Exploring the benefits of discretization of acoustic features for speech emotion recognition. In: Proceedings of 10th conference of the International Speech Communication Association (INTERSPEECH), ISCA, Brighton, UK, pp 328–331
26. Vogt T, André E (2011) An evaluation of emotion units and feature types for real-time speech emotion recognition. This volume
27. Wagner J, Kim J, André E (2005) From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: IEEE international conference on multimedia and Expo, ICME 2005, pp 940–943

[3] http://hcm-lab.de/ssi.html.

28. Wagner J, André E, Jung F (2009) Smart sensor integration: a framework for multimodal emotion recognition in real-time. In: Affective computing and intelligent interaction (ACII 2009), IEEE
29. Wagner J, Jung F, Kim J, André E, Vogt T (2010) The smart sensor integration framework and its application in EU projects. In: Kim J, Karjalainen P (eds) Workshop on bio-inspired human-machine interfaces and healthcare applications (B-Interface 2010), Biostec 2010. INSTICC Press, Valencia, pp 13–21
30. Wobbrock JO, Wilson AD, Li Y (2007) Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Proceedings of the 20th annual ACM symposium on user interface software and technology, UIST '07. ACM, New York, pp 159–168

Johannes Wagner graduated as a Master of Science in Informatics and Multimedia from the University of Augsburg, Germany, in 2007. Afterwards he joined the chair for Human Centered Multimedia of the same University. Among other projects, he has been working on multimodal signal processing in the framework of CALLAS.

Florian Lingenfelser received his M.Sc. degree in Informatics and Multimedia from the University of Augsburg, Germany, in 2009. In 2010 he joined the chair for Human Centered Multimedia of the same University as PhD student. He is currently contributing to multimodal data fusion within the CEEDS project.

Nikolaus Bee works as a research assistant at the lab for Human Centered Multimedia, at Augsburg University. In 2006, Nikolaus Bee won the GALA Award with an application where lifelike characters became aware of users' interest by tracking their eye gaze. His knowledge and research interests involve attentive agents, multimodal interaction and gaze recognition.

Elisabeth André is full professor of Computer Science at Augsburg University and Chair of the Laboratory for Human-Centered Multimedia. In summer 2007 Elisabeth André was nominated Fellow of the Alcatel-Lucent Foundation for Communications Research. In 2010, she was elected a member of the prestigious German Academy of Sciences Leopoldina and the Academy of Europe. Her research interests include affective computing, intelligent multimedia interfaces, and embodied agents.