From R. Cowie and F. de Rosis (Eds.),
Affective Computing and Intelligent Interaction,
***Doctoral Consortium*, Lisbon, Portugal 2006, pp. 1–8**

Real-time Recognition of the Affective User State with Physiological Signals

Florian Hönig, Anton Batliner, and Elmar Nöth*

University of Erlangen-Nuremberg,
Institute of Pattern Recognition (Informatik 5),
Martensstraße 3, 91058 Erlangen, Germany
hoenig@informatik.uni-erlangen.de

Abstract. This study aims at developing methods for recognising the affective user state with physiological signals in near real-time. A multimodal database has been collected in a simulated driving context. Relaxed and stressed states are elicited by giving the participant different tasks. The structured design of the experiment can be used to obtain a preliminary "ground truth"; a fine-grained manual annotation of the perceived stress level is currently being conducted. A data-driven, multi-resolution approach to feature extraction is taken. The classification module can deal with a dynamically varying number of input channels in the case of corrupted signals. For online, user-independent recognition of a relaxed or stressed state during the most clearly defined segments of the experiments, an accuracy of 88.8 % has been obtained using six physiological signals. Current work focuses on a reliable artefact detection, un-supervised user adaption and methods for evaluating the real-time properties of the classification.
Key words: Biosignals, Driving Simulation, Online Stress Recognition

1 Introduction

Research on Human-Computer-Interfaces has recently begun to take into account the affective state of the user. By appropriately reacting to the affective user state, interfaces could not only become more pleasant or entertaining, but also more effective or safer [1]. It is known that affective states have bodily correlates, and thus *physiological signals* could provide the necessary information for acquiring the affective user state. In contrast to other sources of information about the affective user state such as speech or facial expression, most physiological signals are not under voluntary control, and thus cannot be masked up to the same extent. Furthermore, it can be presumed that the signs of affective states in physiological signals are less dependent on individual and contextual factors. Several studies have shown the feasibility of recognising affective user states with the help of physiological signals [2] [3].

A possible application is an in-car infotainment system as for example the one developed in the SmartWeb [4] project. Here, an appropriate reaction to a

* This work was funded by the EC within HUMAINE (IST-2002-507422) and by the German Federal Ministry of Education and Research (BMBF) within SmartWeb (Grant 01 IMD 01 F). The responsibility for the content lies with the authors.

stressed user could be to retain non-vital information in order not to further increase the user's cognitive load. Physiological signals could be integrated, for example a heart rate sensor into the seat or a respiration sensor into the seat-belt.

For a number of conceivable applications, a classification system of affective states would need to work in near real-time and be independent of the user. The present study tries to address some of the principal problems arising from this task. The experimental investigations are done in a simulated car-context; as affective states a relaxed and stressed user state are studied. The key contribution of this study is the investigation of the real-time issues of the extraction and classification of physiological features.

2 Data Collection

In order to apply supervised machine learning algorithms for pattern recognition and for evaluation of classification performance, a labelled database is needed. This poses mainly two problems: the task of obtaining recordings of participants when they are experiencing the different desired affective states, and estimating the *ground truth*, i. e. the actual user state for each point in time. Furthermore, the database must be large enough to appropriately reflect the intra- and interpersonal variability of the physiological signals.

For this work, the DRIVAWORK (Driving under Varying Workload) database has been collected. It is a multi-modal database containing audio, video and physiological recordings of participants in a simulated driving context. Six physiological signals are digitised at 256/2048 Hz with the Mind Media NeXus-10 device: electrocardiogram (ECG), electromyogram at the neck (EMG), skin conductivity between index and middle finger (SC), skin temperature at the little finger (Temp), blood volume pulse at the ring finger (BVP) and abdominal respiration (Resp). The used driving simulation is the *lane change task*[1] (LCT), a PC-based tool for assessing driver distraction which requires only standard consumer equipment [5]. The task of the driver is to switch lanes as quickly as possible according to signs appearing at the roadside; otherwise, the driver is requested to keep the current lane as precisely as possible. As speed is a constant 60 km/h, LCT provides uniform driving demands and readily yields measures of driving performance. The participants drive with only one hand at the steering wheel; the other hand rests on a thigh of the participant and is used for acquiring SC, BVP and Temp in order to keep motion artefacts at a minimum.

The driving is divided into segments of about three minutes duration. During some of these segments, the driver is given additional tasks, intending to elicit an increased stress level: a memory task, question answering and mental arithmetic. All instructions are pre-recorded and given to the participants via headphones. However, the timing is controlled interactively by the experiment conductor which is used in the case of the question and arithmetic tasks to produce an incessant sequence of assignments. At the beginning and end of the experiment, there are two segments of four minutes length without any task; here, the participant is asked to relax. Further segments contain reading and

[1] The authors would like to thank Dr. Stefan Mattes from DaimlerChrysler AG for granting permission to use LCT.

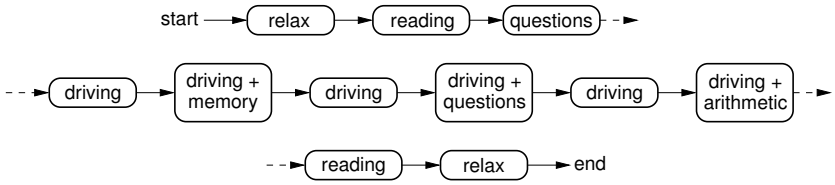

Fig. 1. The structure of the experiment conducted for every participant. The duration of each segment is 3–4 minutes and totals to approximately 35 minutes

question tasks in different contexts. By that, we wanted to elicit speech under different stress levels. The participant is asked to rate each segment on a scale from 1 for very relaxing to 10 for very strenuous. The word "stress" was avoided in the description of this scale as in all text and speech directed to the participant, because conscious reflection about stress might reduce the effectivity of the stress induction. The structure of the whole experiment is depicted in Fig. 1.

The experiment has been performed with 10 female and 14 male participants, yielding 15 hours of usable data. In terms of storage size, the uncompressed physiological data amounts to 1.1 GB, the uncompressed audio data to 9.8 GB, and the video recordings to 216 GB (Sony DV compression). According to the subjective ratings of the participants, the stress induction works well: The average ratings range from 1.5 during the relaxation segments to 7.6 during the mental arithmetic task on top of the driving. The driving segments with additional tasks are perceived more strenuous than those without ($p < 0.001$, Welch two sample t-test). An unexpected result are the relatively high ratings for the first reading and the first question task: they are about as high as the ratings for the questions on top of driving and the reading task directly after the arithmetic task, the presumably most stressful stage of the experiment. That indicates that the elicitation of relaxed speech did not work very well in the experiments.

For the driving segments, the LCT log-files can be used for assessing the effects of the additional tasks on driving performance. The average lane deviation per segment, i. e. the mean distance between a normative model and the actual course of the subject along the track, is noticeably larger when additional tasks are given ($p = 0.04$). However, the overall performance with respect to lane deviation varies strongly between participants; if the average lane deviation is mean-variance-normalised per participant, the difference is marked ($p < 0.001$).

Even more evident effects can be observed for the reaction times for a lane change. The start of a lane change can be detected quite robustly by searching for a maximum resp. minimum of the second derivative of the steering wheel angle within a certain time interval after the lane change instruction on the signs becomes visible. The average reaction time is larger in those segments with additional tasks on top of the driving ($p < 0.001$). During the segment with arithmetic questions, the average reaction time is with 0.76 sec ($\sigma = 0.15$ sec) almost 0.2 sec higher than the average reaction time of 0.57 sec ($\sigma = 0.08$ sec) during the driving segments without additional tasks ($p < 0.001$).

The evidence from the self-assessment and the objective measures obtained from the driving simulation support the conjecture that different stress levels are indeed elicited by the experimental design. Up to a certain extent, that also

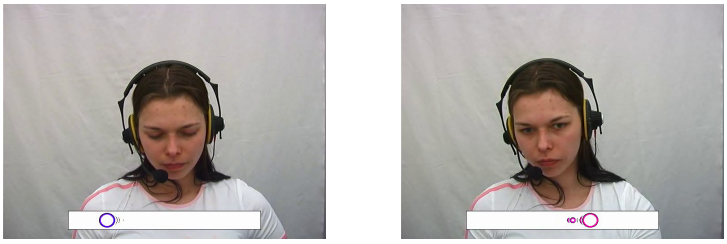

Fig. 2. Snap-shot of the video annotation during a relax segment (left) and the segment with mental arithmetic (right)

justifies using the structured design of the experiment to derive a preliminary ground truth. For example, one can be reasonably sure that the segment with the arithmetic tasks produces a significantly higher stress level than the two relax segments. For the study of real-time recognition of the user state however, this coarse labelling with the help of the segments does not suffice: It does not really represent the dynamic behaviour of the user state and in particular does not allow to infer when transitions between different states take place. Therefore, a fine-grained, continuous labelling of the user state is required. In [3], a continuous stress metric is created from the frequency of objective stress indicators like turning the steering wheel or changing gaze during a real-world driving task. Here, a different approach is taken: Using a setup very similar to the EmoTrace version of the *FeelTrace* tool [6], the subjectively perceived stress level based on the video and audio recordings (both of the participant's utterances and the instructions) is annotated in real-time. The stress level is continuously tracked on a slider that is controlled by the mouse; the cursor has a dragging tail when moved which gives the impression of a slight inertia. The labellers are instructed to assign a value of zero (leftmost cursor position, blue cursor colour) to the maximally relaxed state imaginable for any person and a value of one (rightmost cursor position, red cursor cursor) to the maximally stressed state imaginable for any person. Intermediate states are to be rated according to their similarity to these extreme states; a further clue is given by the middle of the scale (0.5, cursor colour magenta) which marks the boundary between a rather relaxed or a rather stressed state. Figure 2 shows examples from a video annotation session.

In order to get an annotation as consistent as possible, the labellers are shown summary videos of the experiments, giving them a feeling of the stress states to be expected. Before the first experiment is labelled, a video of 8 minutes duration is shown that contains recordings from all 24 participants; before each annotation session a three-minute excerpt from the recording to be labelled is shown. Each experiment is annotated within a single session; the recordings are presented in chunks of 5 minutes duration and 30 seconds overlap followed by a break. Three labellers will annotate the recordings in different order; currently, only one labeller has completed all recordings. For 5 experiments, annotations of two labellers are available; for these, the mean correlation between different labellers is 0.66, the mean absolute difference is 0.16. Figure 3 shows examples of resulting annotations.

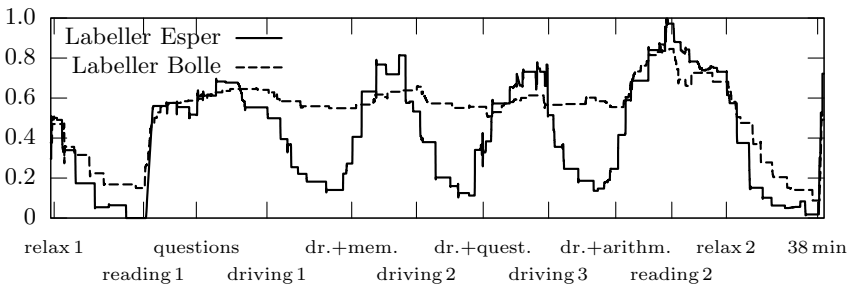

Fig. 3. Annotation of the continuous stress level for a selected experiment. On the x axis, the times when the different segments start are marked out

3 Feature Extraction

The real-time capability required for the classification system has two important consequences: firstly, only samples from the past can be used for analysis of the signals. Secondly, signal processing, feature extraction and classification must be fast enough to process the incoming data in real-time. Our approach [7] tries to address these two issues.

To minimise variability in the signals not related to the affective user state, analysis windows as large as possible would be desirable. On the other hand, large windows containing only data from the past can scarcely capture quick changes and therefore conflict with the goal of real-time classification. Therefore, features are calculated from multiple analysis windows of different lengths: 1, 5, 20 and 60 seconds. This multi-resolution approach aims at combining the stability of large analysis windows with the capability of small windows to reflect quick changes.

The different physiological signals each exhibit individual properties. This makes the engineering of a dedicated set of features necessary for each physiological modality and each application-dependent set of states. In our approach, this is addressed by first computing a large number of multi-purpose features such as mean value, standard deviation or slope for each analysis window. Then, the labelled dataset is utilised to create task- and signal-specific features by means of a data-driven transform, the Fisher linear discriminant analysis (LDA).

Two different feature sets are provided. The *moving features* are computed recursively for each new sample and thus have a constant computational complexity with respect to the length of the analysis window and the desired frequency of feature vector computation. A ring-buffer is used to store the necessary sample history. In effect, these features can be computed very quickly for all window sizes and are well-suited for a possible implementation on limited hardware. The recursive calculation is illustrated by the update rule for the mean value μ_n of a window containing w samples at the n-th sample x_n: $\mu_n = \mu_{n-1} - x_{n-w}/w + x_n/w$. If floating point numbers are used, and unless w is small, errors due the numerical instability of adding and subtracting small values accumulate and render the result useless with time. This can be solved by periodically providing a mean value calculated anew; substituting the recursively calculated value every w samples results in a reasonable degree of numerical sta-

bility and only increases the computational effort by a constant factor of about 2. With similar techniques, also mean values as would result from a triangle- and bell-shaped window can be computed recursively. Further recursively computed features are e. g. the slope of the regression line, a smoothed derivative, the variance, mean absolute rise, fall and change and approximations of the minimum, maximum, median and the amplitude. Furthermore, the square, the square root or the absolute value of some features is added, e. g. the absolute value of the slope or the square root of the variance, yielding the standard deviation. In total, 50 moving features are calculated for each analysis window.

The *sliding features* drop the need for a sample history, thus resulting in a memory requirement independent of the window length. This is favourable for a possible implementation on hardware with small memory. The recursive calculation again is illustrated by the update rule for the mean value $\mu_{\alpha,n}$ and a parameter $\alpha < 1$: $\mu_{\alpha,n} = \alpha \cdot \mu_{\alpha,n-1} + (1-\alpha)x_n = (1-\alpha)\sum_{i=0}^{\infty} \alpha^i x_{n-i}$, i. e. $\mu_{\alpha,n}$ is the mean value of the signal multiplied with an exponentially decaying window function. Depending on the desired (nominal) length w of the analysis window, α is chosen such that the standard deviation of the exponential window corresponds to the standard deviation of a rectangular window of length w: $\alpha = 1 - 2\sqrt{3}/w$. Due to the fact that the window function never actually reaches zero, large outlier values of a signal can corrupt the mean value for a long time. Therefore, $\mu_{\alpha,n}$ is periodically substituted by a value that would result if the window function was set to zero after 99 % of its mass. Again, the computational effort is only increased by a constant factor of about 2. With similar techniques, and by carefully combining mean values calculated with slightly different values for α, 44 "sliding" equivalents of most of the moving features can be calculated, including a sliding version of the smoothed derivative and the regression slope.

Before computing the features, four additional signals are derived from the actually recorded signals: The heart rate from the ECG channel (HR-ECG), the heart rate from the BVP channel (HR-BVP), the lag between ECG and BVP (Lag) which can be regarded as a surrogate parameter of the systolic blood pressure and the respiration rate (Resp-rate). The reason for treating those as separate signals will be given in the next section.

4 Classification

Physiological signals are prone to artefacts caused by motion, pressure and other disturbances. This is especially true if they are to be applied in a real-world application and not just in the laboratory under carefully controlled conditions. If a signal becomes predominantly corrupted for some interval of time, the classification module should be able to ignore the respective channel.

Our approach can deal with a variable number of input channels. First, it is decided for each channel whether the signal is corrupted. Currently, this artefact detection is only a simple rule disqualifying signals with unplugged sensors or physically implausible values for the derived signals HR-ECG, HR-BVP, Lag and Resp-rate. All signals marked as corrupted are excluded from further processing for the current point in time. Now a reason for treating the derived signals separately can be given: if a derived signal is corrupted, a base signal such as BVP might still be of value and should not automatically be excluded.

For each remaining channel, the feature vector resulting from the LDA transformation is scored separately with a Gaussian mixture model consisting of 10 mixture components. The resulting probabilities are, assuming statistical independence between the different physiological signals, combined by multiplication, yielding a final score for each class. The approach has been evaluated for discriminating between a relaxed and stressed user state using a subset of the DRIVAWORK database: the classification accuracy is evaluated only during the two relax segments and the segment with the arithmetic task. This amounts to 4.3 hours of data. It is assumed that the affective state of the person is the one intended by the experimental design. Classification is done with a frequency of 1 Hz, so the number of used feature vectors is about 15600. Note that the chosen online classification task is more difficult than the task of discriminating previously defined, relatively large segments in an offline manner as studied e. g. in [3] in the following sense: the context of 60 seconds available to the classification module is relatively small, in addition, it is only taken from the past. So, a considerable fraction (28 %) of the feature vectors is computed from intervals that are not completely contained within the relax and arithmetic segments. However, the task is still artificially simplified by the fact that the studied segments are well separated. An evaluation of the whole experiment has not yet been done because of the difficulty of defining a ground truth for the other segments.

All evaluations are done using person-independent 10-fold cross-validation, i. e. each pair of train and test set is disjoint with respect to the participants. The class-wise averaged recognition rates are reported. Using the individual signals alone, recognition rates between 49.2 % and 84.3 % resulted. When all signals were combined for classification, an accuracy of 88.8 % was achieved. When doing an offline evaluation with centred analysis windows, the accuracy rose slightly to 89.3 %. Simulating user adaption by normalising the mean and variance of all features per participant (before estimating the LDA transform), recognition rates of 92.0 % and 93.6 % are reached for the online and offline case, respectively. Given the difficulty of the task of online, person-independent classification these results are quite satisfactory and prove that the data-driven, multi-resolution approach works well. Some of the success however, may be attributed to external factors such as the fact that the participants speak and operate the steering wheel during the stress segment even though special care has been taken to prevent motion artefacts. Our structured design allows to examine this issue more closely.

5 Future Work and Conclusion

The classification of the user state was done online, i. e. only using data from the past. However, the recognition rates do not contain information about the real-time capabilities of the classification system in terms of a reaction time to changes of the user state. Also, the coarse annotation derived from the structured design of the experiment does not allow to evaluate such properties. However, the fine-grained manual annotation of the stress level that is currently being conducted could provide answers to these questions. For assessing the delay of the classification system, the sequence of predicted values could be aligned to the reference values. Using this technique, a preliminary analysis using linear

regression to predict the annotated stress level from the calculated features indicated that on average, the classification system has indeed a low delay in reaction speed to user state changes. However, this needs to be examined in more detail.

Further research will include a more sophisticated artefact detection module. A possibility could be using the above classification system to decide whether the signal is corrupted on the basis of the calculated features; as a reference, those sections of the signal could serve where user state classification fails. Other issues will be the development of recursively calculated spectral features and a method for un-supervised user adaption in order to increase classification accuracy. Finally, real-life recordings of driving school lessons are planned.

The main contribution of this study is the implementation of the real-time recognition of the user state based on physiological signals. A number of problems arising from this task have been addressed, and new solutions have been proposed: a signal analysis scheme that combines the stability of large analysis windows with the capability of small windows to reflect quick changes; a large number of fast, generic features that are transformed into a task- and signal-specific reduced feature vector using a data-driven transform; a classification module that can handle a dynamically varying number of channels in the case of signal corruption. To test the approaches, a database containing 15 hours of recordings from 24 participants in different stress levels has been collected; the effectiveness of stress elicitation has been confirmed by objective measures. Methods for defining a ground truth have been provided, including a fine-grained manual annotation of the perceived stress level. The evaluation showed that the presented approaches are well suited for the task of online, user-independent stress recognition. Finally, areas for further research were identified: evaluation of the reaction time of the classification system, sophisticated artefact detection, recursively calculated spectral features and un-supervised user adaption.

References

1. Nass, C., Jonsson, I.M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L.: Improving automotive safety by pairing driver emotion and car voice emotion. In: CHI '05 extended abstracts on Human factors in computing systems, New York, NY, USA, ACM Press (2005) 1973–1976
2. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(10) (2001) 1175–1191
3. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems **6**(2) (2005) 156–166
4. Reithinger, N., Herzog, G., Blocher, A.: Smartweb - mobile broadband access to the semantic web. KI Zeitschrift (2) (2007) 30–33
5. Mattes, S.: The lane-change-task as a tool for driver distraction evaluation. In: Quality of Work and Products in Enterprises of the Future, Annual Spring Conference of the GfA/17th Annual Conference of the ISOES. (2003) 57–60
6. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: An instrument for recording perceived emotion in real time (2000)
7. Hönig, F., Batliner, A., Nöth, E.: Fast recursive data-driven multi-resolution feature extraction for physiological signal classification. In: 3rd Russian-Bavarian Conference on Biomedical Engineering, Erlangen (2007)