Predicting Continuous Stress Ratings Of Multiple Labellers From Physiological Signals

Hönig F, Batliner A, Eskofier B, Nöth E
Institute of Pattern Recognition, University of Erlangen-Nuremberg
hoenig@informatik.uni-erlangen.de

Abstract. In this paper, we study means of estimating a person's current stress level from physiological signals. Generic, data-driven features are extracted from multiple channels and used to predict a continuous stress level with regression techniques. Allowing for temporary signal corruption by artefacts, our approach can handle a variable number of input channels. Additionally, methods for estimating the reaction time of the system are proposed. The evaluation of the approach with reference annotations of three labellers shows promising results.

1 Introduction

The aim of this study is to evaluate methods for estimating a person's current stress level from physiological signals in real time. We use the Drivawork database which contains recordings of six physiological signals, audio and video during varying levels of workload in a simulated car-drive (see [1] for details). It contains 15 hours of physiological recordings from 24 participants. Relaxed and stressed states have been elicited by giving the participant different tasks, partly on top of the driving task. Subjective and objective measures support the effectiveness of this approach. In prior studies, we have used the structured design of the recording experiment to derive stress labels for whole segments; the two classes stress/non-stress could be predicted person-independently with relatively high reliability (86-94 % depending on the chosen evaluation subset) even when using only sensor data from the past 60 seconds. However, when wishing to study the real-time properties of such a classification system, more specifically the reaction time to user state changes, such coarse labels will not do. In [2] 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 to derive a continuous stress metric. We take another approach here, and use the manual ratings of three labellers available in the Drivawork database. These labels have been created by tracing the perceived stress level of the participant on a slider while watching the video recording (audio included) of the experiments. The position of the slider is mapped to a value between 0 for a maximally relaxed and 1 for a maximally stressed state. The ratings are read out once per video frame; thus, these labels have the capability to reflect even quick user state changes. The ratings of two labellers for one participant have a Pearson correlation coefficient of 0.76 and an absolute deviation of 0.13 on average. Present attempts to translate the continuous labels into discrete classes did not yield acceptable agreement rates among the three labellers; therefore, we use the continuous ratings directly and predict them with regression techniques.

2 Methods

Six physiological signals are used in this study: electrocardiogram (ECG), electromyogram measured at the neck, skin conductivity between index and middle finger, blood volume pulse (BVP) at the ring finger, skin temperature at the little finger and abdominal respiration. From these signals, three derived signals are created: heart rate from ECG and from BVP and the lag between ECG and BVP which can be regarded as a surrogate parameter of the blood pressure. To be robust against sensor failure or corruption by artefacts, each channel is first analysed separately; only valid channels are then combined for the final result; cable disconnects and physically implausible values in the derived signals will lead to the temporary exclusion of a channel.

For each signal s, features are computed from multiple analysis windows (1, 5, 20 and 60 seconds) to capture quick changes as well as context for the analysis. As we aim for online stress estimation, these analysis windows are causal, i.e. they contain only data from the past. From each analysis window, a relatively large number of features like mean, standard deviation or slope are calculated. For the present study the 50 efficient “moving features” as described in [3] are used. The 200 features from all four analysis windows are mean-variance-normalised and then transformed to a reduced vector of dimension 100 with Principal Component Analysis (PCA). Then, linear regression is applied to compute an estimate $\hat{y}_s$ of the continuous stress rating y. Although reference stress values are given with a frequency equal to the frame rate of the video recording, feature vectors are only computed with a frequency of 4 Hz; the reference values are down-sampled accordingly. Two methods for combining the variable number of predicted ratings from the currently valid channels are studied: first, linear estimation; second, linear regression again.

Linear estimation is a weighted averaging indirectly proportional to the mean squared error of the respective input; it yields the minimal squared prediction error if the inputs are statistically independent and unbiased [4]:

$$\hat{y}_{\text{est}} = \Big(\sum_{s \text{ valid}} w_s \hat{y}_s\Big) / \Big(\sum_{s \text{ valid}} w_s\Big), \qquad w_s = 1 / \Big(\sum_{i \in \text{train}} (\hat{y}_s^{(i)} - y^{(i)})^2\Big).$$

For linear regression, the transformation parameters depend on the subset of currently valid input channels; however, they can be calculated on the fly with a computational effort cubic in the number of inputs but independent of the number of training vectors—feasible for a moderate number of input channels as in our case (for cases with many inputs more efficient approaches as described in [5] could be used). Let

$$\hat{y}_{\text{reg}} = \boldsymbol{t}^{\mathrm{T}}(\hat{\boldsymbol{y}} - \overline{\hat{\boldsymbol{y}}}) + \overline{y}, \quad \hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_S)^{\mathrm{T}}$$

be the linear regression if all S channels are valid. Then $\boldsymbol{t}$ is determined from N uncorrupted training samples by

$$\boldsymbol{t}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{b}^{\mathrm{T}};\ \boldsymbol{A} = \hat{\boldsymbol{Y}}\hat{\boldsymbol{Y}}^{\mathrm{T}},\ \boldsymbol{b}^{\mathrm{T}} = \boldsymbol{y}^{\mathrm{T}}\hat{\boldsymbol{Y}}^{\mathrm{T}},\ \hat{\boldsymbol{Y}} = (\hat{\boldsymbol{y}}^{(1)}, \ldots, \hat{\boldsymbol{y}}^{(N)}),\ \boldsymbol{y}^{\mathrm{T}} = (y^{(1)}, \ldots, y^{(N)}).$$

$\boldsymbol{A}$ and $\boldsymbol{b}$ are computed offline; during operation, depending on the currently valid channels, the deletion matrix respectively vectors

$$\boldsymbol{A}' = (a_{ij})_{i,j \text{ valid}},\ \boldsymbol{b}' = (b_i)_{i \text{ valid}},\ \hat{\boldsymbol{y}}' = (\hat{y}_i)_{i \text{ valid}}$$

are formed and $\boldsymbol{t}'$ and $\hat{y}_{\text{reg}}$ are computed from

$$\hat{y}_{\text{reg}} = \boldsymbol{t}'^{\mathrm{T}}(\hat{\boldsymbol{y}}' - \overline{\hat{\boldsymbol{y}}'}) + \overline{y}, \quad \boldsymbol{t}'^{\mathrm{T}}\boldsymbol{A}' = \boldsymbol{b}'^{\mathrm{T}}.$$

Figure 1 shows an example for each method. It can be seen that linear estimation tends to produce a “shrunken” prediction (values near the mean) while linear regression produces noisy estimates from the variable-dimensional input.

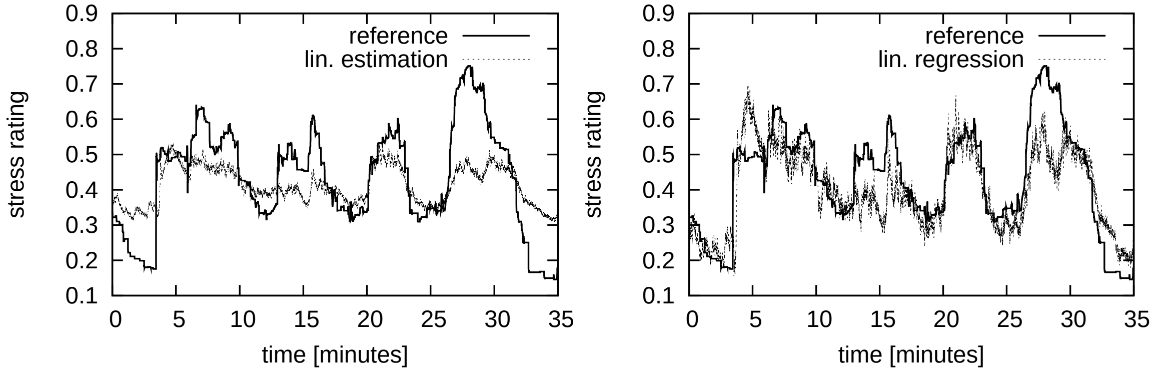

Fig 1. Predicting the continuous stress level from physiological features. The variable number of input channels is handled by linear estimation (left) or accordingly re-computed linear regression (right). The displayed sequence was not used for parameter estimation.

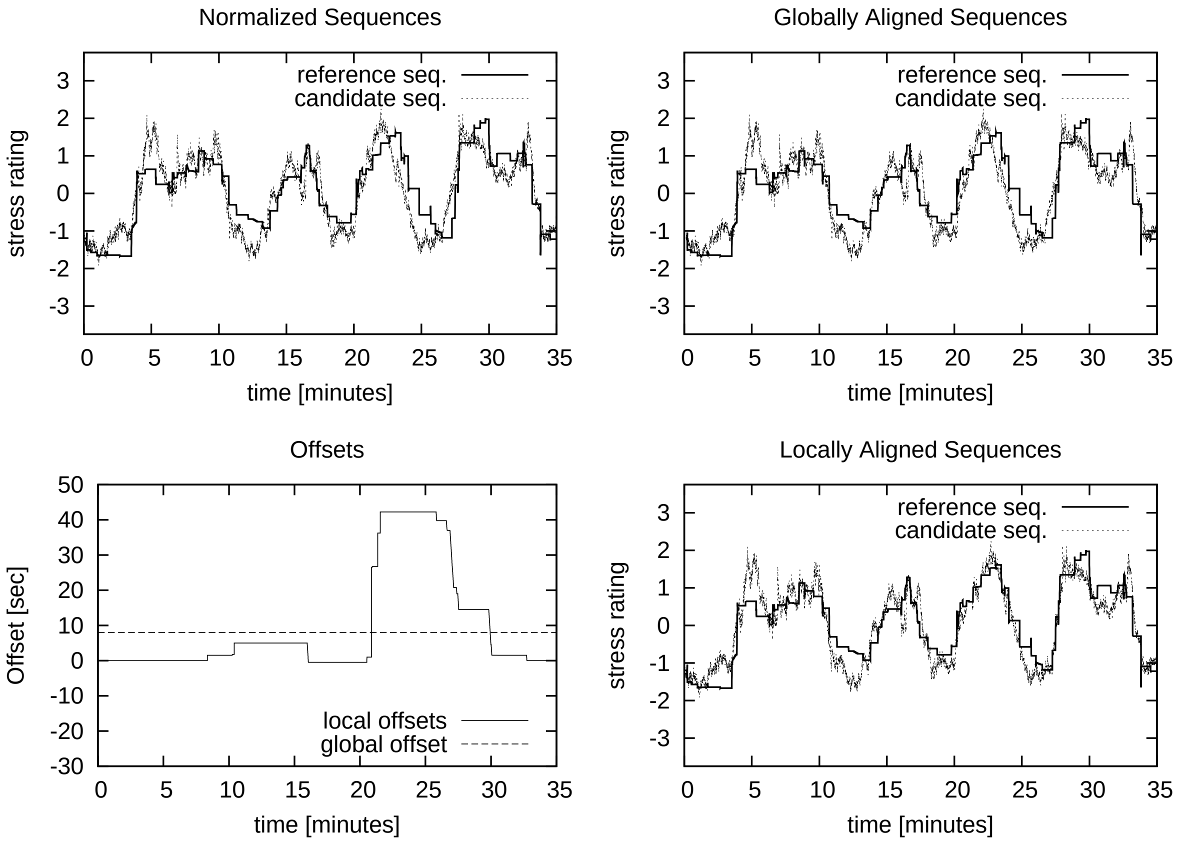

Fig 2. Aligning reference and candidate sequence in order to get an estimate for the reaction time of the system. Offset is the time shift introduced to the candidate sequence; thus it equals the negative reaction time. In this example, global alignment determines the candidate sequence to be 8 sec. ahead of the reference; local alignment yields 8.9 sec. on average.

For estimating the reaction time of the system, the sequence of predicted stress ratings, i.e. the *candidate sequence*, is aligned to the reference sequence. We compare both a global alignment, estimating the reaction time by determining a constant time shift between candidate and reference sequence, and a local alignment from dynamic time warping, using the mean time shift as an estimate for the reaction time. In both cases, the sequences to be compared are first mean-variance-normalised. The global alignment computes the shift that maximises the correlation between reference and alignment. The correlation values are computed efficiently with the help of the Fast Fourier Transform; the values are normalised by the number of summation coefficients within a valid range to prevent a bias in favour of small shifts. That again causes high correlation values for large shifts; restricting the maximisation to an interval of ± 2 minutes yields meaningful results. The local alignment is determined as the assignment path that minimises a combined cost made up from the total absolute difference between assigned candidate and reference values and the number of insertions and deletions contained in the path. The cost of an insertion or deletion is chosen equal to the mean absolute difference between the original (i.e. not aligned) candidate and reference sequence. Figure 2 illustrates the two methods using an example.

3 Results

All evaluations were done on a subset of the Drivawork database: the segments during which the participants had to drive (totalling to 7.8 hours). Results from a 10-fold person-independent cross-validation are reported, i.e. each pair of train and test set was disjoint with respect to participants. Table 1 lists the mean Pearson correlation coefficient and absolute difference between reference and candidate sequence using linear estimation and regression for various setups: *L1*, *L2* and *L3* refer to the ratings of the individual labellers; *mean* to the average rating of all three labellers and *mean 2 ∅* to the average correlation coefficients when taking the mean of only two labellers as reference. The complete multi-resolution feature set described above is termed *online multi*; *online single* uses only the features from the 60-second analysis window. The *offline* feature configurations use non-causal, centred analysis windows. Lastly, *offl. m. norm* refers to the *offline multi* features mean-variance-normalized per participant. For comparison, results with the trivial estimator that puts out the mean target value of the training set are also given. In almost all cases, linear estimation gives a higher correlation than linear regression. Only for the absolute difference, linear regression is sometimes better. In all cases, multiple resolutions give superior results when compared to a single analysis window. Furthermore, it is obvious that averaging the reference over multiple labellers improves results. For interpreting the precision of the system, the ratings of each labeller were compared to the averaged ratings of the other two labellers (for the car drive subset). Here, a mean correlation coefficient of 0.761 (average absolute difference: 0.129) resulted. When comparing these figures to the appropriate figures of the system (*mean 2 ∅*), it becomes obvious that the system performs moderately well but does not reach the performance of the labellers.

For the estimated reaction time, a less clear picture resulted. While estimating reaction times between the labellers (see Table 2) gave relatively consistent figures, results for stress values predicted with physiological features (see Table 3) are often contradictory. For example, using global alignment, an estimated reaction time of -4.6 seconds (i.e. candidate sequence is ahead of reference) resulted while local alignment gave +2.1 seconds (features: *online multi*, reference: *mean*). Also, the estimates show a large standard deviation across participants. However, the fact that the *L2* labels tend to be delayed compared to those of the other labellers seems to be reflected in the results for the physiologically predicted stress levels: The automatic system has in all cases a lower reaction time when trained and tested with the *L2* reference than when using the other labellers. Multi-resolution analysis did not consistently show a lower reaction time, however, comparability is questionable here due to the differing precision achieved with the *multi* and *single* features.

Tab 1. Average correlation between reference and candidate sequence for various setups. The mean absolute difference is given in parentheses.

Method	**Labeller** *Features*	**L1**	**L2**	**L3**	**mean**	**mean 2 ∅**
Trivial	*none*	**.000** (.168)	**.000** (.070)	**.000** (.106)	**.000** (.099)	**.000** (.103)
Linear Estimation	*online multi*	**.637** (.170)	**.506** (.084)	**.642** (.082)	**.688** (.093)	**.664** (.098)
	online single	**.608** (.175)	**.465** (.084)	**.604** (.084)	**.646** (.096)	**.625** (.101)
	offline multi	**.629** (.172)	**.495** (.083)	**.651** (.082)	**.680** (.093)	**.658** (.098)
	offline single	**.612** (.176)	**.458** (.085)	**.619** (.084)	**.651** (.096)	**.628** (.101)
	offl. m. norm	**.669** (.160)	**.589** (.075)	**.674** (.077)	**.717** (.085)	**.698** (.090)
Linear Regression	*online multi*	**.612** (.180)	**.503** (.091)	**.621** (.082)	**.671** (.095)	**.647** (.101)
	online single	**.582** (.186)	**.470** (.089)	**.569** (.086)	**.625** (.099)	**.602** (.105)
	offline multi	**.604** (.179)	**.489** (.093)	**.638** (.078)	**.659** (.093)	**.640** (.100)
	offline single	**.601** (.183)	**.455** (.095)	**.606** (.082)	**.648** (.097)	**.623** (.103)
	offl. m. norm	**.660** (.151)	**.538** (.080)	**.657** (.075)	**.704** (.081)	**.680** (.087)

Tab 2. Estimated average reaction time in seconds between stress labels of different labellers. The standard deviation across participants is given in parentheses.

Alignment	**Reference** *Candidate*	**L1**	**L2**	**L3**
Global	*L2*	**3.8** (6.7)	-	-
	L3	**0.8** (7.7)	**-3.7** (7.2)	-
	mean (L1, L2)	-	-	**1.0** (6.7)
	mean (L1, L3)	-	**-2.2** (4.6)	-
	mean (L2, L3)	**3.0** (5.3)	-	-
Local	*L2*	**12.9** (19.6)	-	-
	L3	**5.8** (11.2)	**-6.8** (9.2)	-
	mean (L1, L2)	-	-	**-4.3** (9.5)
	mean (L1, L3)	-	**-13.0** (14.3)	-
	mean (L2, L3)	**6.6** (11.2)	-	-

Tab 3. Estimated average reaction time in seconds of the stress level predicted with linear estimation. The standard deviation across participants is given in parentheses.

Alignment	**Labeller** *Features*	**L1**	**L2**	**L3**	**mean**
Global	*online multi*	**-6.0** (12.4)	**-9.4** (14.4)	**-2.3** (27.1)	**-4.6** (9.4)
	online single	**-10.2** (25.7)	**-17.3** (29.1)	**3.4** (25.0)	**-8.6** (24.1)
	offline multi	**-4.1** (24.3)	**-14.1** (20.2)	**-1.7** (19.6)	**-5.0** (20.8)
	offline single	**-8.8** (15.5)	**-21.7** (35.5)	**-4.1** (13.0)	**-6.8** (13.2)
Local	*online multi*	**2.0** (17.8)	**0.0** (11.6)	**0.4** (15.6)	**2.1** (15.4)
	online single	**3.7** (20.2)	**-3.6** (16.3)	**3.8** (16.6)	**0.2** (15.1)
	offline multi	**-0.9** (22.6)	**-3.1** (13.1)	**-2.3** (19.3)	**-3.3** (18.3)
	offline single	**-3.5** (17.3)	**-5.8** (16.4)	**-2.2** (15.0)	**-3.4** (15.9)

4 Discussion

The results show that the generic feature extraction approach developed in [3] is also applicable to online regression; the superiority of the multi-resolution analysis approach (compared to a single analysis window) was confirmed here also. The fact that averaging over multiple labellers increases precision indicates that the manual labelling procedure yields a noisy but meaningful reference. An advantage of the multi-resolution analysis in terms of reaction time has not been proven so far; the results for the estimated reaction times are partly contradictory. An explanation might be that the alignment cannot handle well the noise present in the estimated stress level sequences.

Future work will concentrate on additional features to increase precision and on developing better estimates of the reaction time. More advanced artefact recognition will be developed to fully exploit the ability of the approach to deal with a variable number of input channels.

References

[1] Hönig F, Batliner A, Nöth E. Real-time recognition of the affective user state with physiological signals. Affective Computing and Intelligent Interaction (ACII 2007), Doctoral Consortium: 2007:1-8.

[2] Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 2005;6(2):156-166.

[3] Hönig F, Batliner A, Nöth E. Fast recursive data-driven multi-resolution feature extraction for physiological signal classification. 3rd Russian-Bavarian Conference on Biomedical Engineering (RBC 2007): 2007:47-52.

[4] Rubin DB, Weisberg S. The variance of a linear combination of independent estimators using estimated weights. Biometrika 1975 62(3):708-709.

[5] Miller A, Subset selection in regression, 2nd edition, Chapman & Hall 2003.