3E3-4

A Hierarchical Model of Authentic Work Stress Using Physiological Signals and Stress Coping Profiles

Juan Lorenzo Hagad*1 Ken-ichi Fukui*2 Masayuki Numao*3

Institute of Scientific and Industrial Research, Osaka University, Japan

Utilizing data from various sources to build multimodal models has been shown to be an effective way to build more accurate and flexible models of mental stress. However, traditional machine learning techniques lack the ability to effectively identify salient inter-modal correlations when diverse modalities are use together. In this work we investigate the efficacy of multimodal models of stress built using a combination of psychological and physiological data. A monitoring platform and unobtrusive wearable sensors were used to gathered data from subjects engaged in authentic work activities. Models were built by combining psychological data from stress coping profiles and physiological signals from the sensors then using self-annotated stress annotations to establish ground truths. A performance comparison was then made between standard machine learning approaches and deep multimodal learning. The results indicate that significant improvements can be achieved by applying deep multimodal feature learning to construct mental stress models.

1. Introduction

Stress can be defined as a biochemical or physiological change in response to internal and external stressors. It is recognized by clinical studies [3] as a risk factor for a number of cardio-vascular diseases and is one of the leading causes of work disabilities worldwide. Due to its pervasiveness, automated mental stress monitoring and diagnosis has gained popularity in recent years and is proving to be a potential key technology for addressing more severe mental health issues such as depression.

1.1 Multimodal Stress Models

Multimodal models involve using data from various sources. In stress monitoring, this technology has seen major advancements from the emergence of modern sensors. Majority of the existing works rely primarily on physiological and environmental signals. Of these, the most commonly used are galvanic skin response (GSR)[7, 6, 10] and heart rate (HR)[7, 6, 10].

In [10] they used galvanic skin response (GSR), blood volume pulse, pupil diameter and skin temperature to detect changes in stress levels. Models were built to distinguish between two states: stressed and relaxed, and the study achieved over 90% accuracy using support vector machines (SVM). Other studies such as [7] followed a similar data gathering framework. In [7] they tested out-of-laboratory experiments using the Intel Shimmer platform which includes ECG, GSR and accelerometers. In the aforementioned work, stress was induced using the Stroop test and mental arithmetic. Using accelerometer data, they showed that additional activity context can also be used to improve detection performance. In these, and many other works, it is common to use traditional machine learning methods such as SVM and neural networks (NN) and train these on simple combinations of the multimodal data. In this work, we will show that these approaches are lacking. Aside from this, many studies tend to rely on data from artificial stress inducing tests such as the Stroop Test and other simulated challenges. While these are designed to stimulate authentic stress responses, there is a tendency towards exaggerated conditions and stress responses. As a result, the resulting model may not properly represent the full spectrum of subtle stress responses that one may encounter in real-world scenarios. On the other hand, using naturalistic data has its own challenges since samples from natural environments are susceptible to noise and have a tendency to feature less pronounced expressions of stress. Furthermore, there is the challenge of gathering ground truth labels. However, it is necessary to investigate such models to discover authentic features and build an appropriate model of real-world mental stress.

To offset the effects of noise and other variabilities, we applied deep multimodal learning to discover useful patterns within each modality and between modalities. We also attempt to merge data from psychology and physiology by combining coping profiles with wearable sensor signals, respectively. It has long been recognized through evidence from medicine and psychology that the stress response is not simply a function of the severity of the stressor, but is also borne from the ability of the organism to cope with it [9]. Thus, individual coping variables can be used as factors that affect the stress response.

2. Methods

2.1 Data and Annotations

All experiments were performed on computers equipped with our purpose-built monitoring and annotation software. Each subject conducted self-regulated work activities on their personal PCs while using the software. This allowed the experimenter to record the subjects personal profiles

Contact: Juan Lorenzo Hagad, Department of Architecture for Intelligence, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan, +81-6-6879-8426, hagad@ai.sanken.osaka-u.ac.jp

and work activities without direct supervision. Since most experimental procedures involved minimal interaction between the experimenter and the subjects, transmission of experimenter biases was also minimized.

Self-reported stress annotations were made immediately following the hour-long work sessions. Using webcam and desktop recordings, subjects reviewed and selected time segments in the recorded video where they performed various work tasks. For each task, they identified the amount of stress they felt at the time. A selection of work task categories and stressors were provided by the UI, and a 4-point Likert-scale (1=very low, 4=very high) was used for the stress annotations.

2.2 Physiological Signals

Participant physiological signals were measured using wearable ECG sensors and wrist sensors. These wireless, wearable devices allowed continuous measurement of heart rate (HR) and skin conductance (SC), respectively.

Heart rate variability (HRV) features were extracted from the HR data. These are HR features that have been shown to have strong correlations with autonomic stress responses [4]. Specifically, the following features were used: Average of NN intervals (AVNN), Standard deviation of NN intervals (SDNN), root-mean-squared differences between adjacent NN intervals (rMSSD), percentage of differences between adjacent NN intervals greater than 50ms (pNN50), spectral power measures of NN intervals of varying frequencies (ULF, VLF, LF, HF) and the ratio of low to high frequency power (LF/HF).

For GSR, two major components of the conductance signal were analyzed: skin conductance level (SCL) and the skin conductance response (SCR) [1]. These features cover different aspects of sympathetic neuronal activity. SCL reflects the *tonic* level or the slowly changing component of the GSR signal, while SCR or rapid *phasic* refers to the faster changing elements of the signal. These measurements were obtained by using software included with the wearable devices. The values were then cleaned and time-dependent statistics were extracted such as the mean, variance, and difference from the baseline levels.

2.3 Psychological Profiles

To build the personal profile, subjects answered the COPE Inventory [2], a questionnaire designed to assess coping responses in response to stressful situations. It determines a person's inclination towards exhibiting responses that are expected to be either functional or dysfunctional. Those with dysfunctional coping mechanisms are expected to be more prone to the negative effects of stress. By including these factors into our machine-learned models, we hope to be able to reduce the ambiguity of stress features relative to each subject.

2.4 Baseline Machine-Learners

Supervised and unsupervised machine learning techniques were used to build baseline stress models to compare with the deep learning models. These were selected from machine learning techniques most commonly featured in related works. For the supervised models, we used support vector machines (SVM) and mutilayer perceptrons (MLP), and for the unsupervised model we used k-means clustering. The SVMs featured used radial basis function (RBF) kernels since these have been shown to be highly flexible for classification tasks. The MLPs featured a single hidden layer with a number of nodes equal to half of the sum of the number of input attributes plus the number of output classes. Finally, the number of k-means clusters were adjusted to match the number of output classes.

2.5 Deep Feature Learning

In this work we implemented Deep Learning using Autoencoders [8]. These artificial neural network structures are similar to traditional feed-forward networks. In its most basic form it includes input, hidden, and output layers. Unlike typical neural networks, autoencoders learn the ideal parameters to generate an output that is a reconstruction of the inputs. Through this process, the hidden layers are able to discover latent features that can efficiently represent the training data. Specifically, we applied Denoising Autoencoders (DA) a form of autoencoder that is trained by reconstructing using stochastically corrupted versions of the input.

The deep learning structures in this work used autoencoders stacked in a greedy layerwise fashion to form a deep network similar to stacked Restricted Boltzmann Machines (RBM) in deep belief networks [8]. The different levels of the stack allow learning multiple levels of abstraction and can be used to learn inter-modal features. However, strong intra-modal feature correlations may still prevent the discovery of some important inter-modal features.

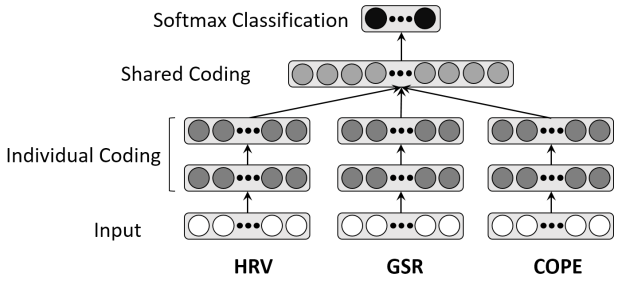

Figure 1: Multimodal Deep Learning Structure

In [5], an study was performed to measure the saliency of multimodal features discovered by different deep RBM structures. It was discovered that models trained on shallow concatenations of audio and video features were not able to capture effective correlations across the modalities. Bimodal deep RBMs, those that featured pretraining at different modal levels, were much better able to capture these correlations. In this work, we apply the same concept, however using stacked autoencoders to build the structure shown in Figure 1.

3. Experiments and Results

Data was collected from 4 healthy male participants aged 20-32. All subjects were graduate school students from Os-

aka University. Each subject performed at least 5 work sessions with each session lasting 1 hour. These annotations resulted in 184 usable work task segments labelled with stress levels from 1 (very low stress) to 4 (very high stress). Each task segment lasted around 5 to 30 minutes. HR was recorded at a sampling frequency of 256Hz while SC was recorded at 128Hz. All frequency-domain and time-domain features were extracted over task segments.

3.1 Baseline Results

The baseline performance for the standard machine models are shown in Table 1. Classification performance was measured using stratified 10-fold cross-validation accuracy. Based on these results, all models performed better than random classification with MLPs showing the best performance.

Table 1: Baseline Performance Results

SVM	MLP	K-means	Random
42.069%	47.414%	40.702%	29.501%

3.2 Denoising Autoencoder Results

For the first round of experiments, we built and tested models using Denoising Autoencoders (DA). For the following experiments we used DAs with 5x overcomplete hidden units for the combination of physiological features (105 units) connected to a logistic regression layer. Pretraining was performed over 100 epochs and with a 0.001 learning rate. The per-fold results are listed in Table 2. When comparing the performance of DAs to MLPs, we noted a rise in mean accuracy from 47.90% to 50.53%. However, statistical analysis via a paired t-test indicated that this was not sufficiently significant (p=0.16).

3.3 Stacked Autoencoder Results

In the next investigation, we attempted to discover features through a deep structure. We built and tested Stacked Denoising Autoencoders (SDA) and compared their performance with the previous DAs on the 4-class dataset. For the SDA we used 3 hidden layers based on results cited in [8]. Once again, we used 5x overcomplete hidden units (105 units) for each of the hidden layers. Pretraining was performed using 100 epochs and at a 0.001 learning rate.

Referring to the results in Table 2 and comparing the results of the single layer DA and the 3-layer SDA, there is a slight reduction in performance. Intuition states that adding more layers may eventually lead to improvements, so to confirm that these were optimal results we also tested models with additional layers. As shown by the pattern of performance in Figure 2, deviating from 3 layers actually leads to similar or worse performance. Error bars indicate standard error over 10-fold cross-validation accuracy. Basically, results indicate that using 1 or 2 layers leads to a high variation in performance, while adding more layers beyond 3 leads to a dramatic decrease in performance.

3.4 Multimodal Deep Learning Results

For the final experiment we modified our approach by applying multimodal deep learning (MDL) [5]. Each modality

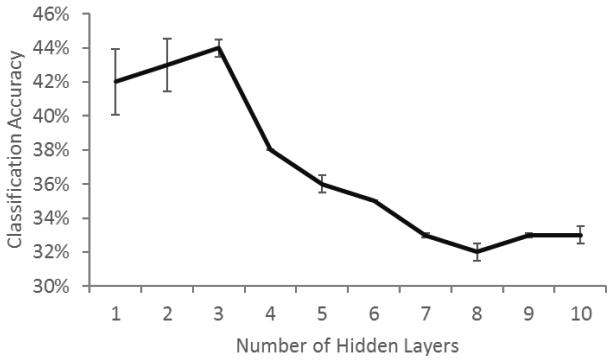

Figure 2: Stack Depth Test Results

was pre-trained as its own isolated denoising autoencoder. This allowed the discovery of unimodal latent features. On top of these, we placed a fully-connected autoencoder layer intended to learn a shared coding for the inter-modal features. Finally, the last layer was a softmax logistic regression layer used for supervised learning and classification.

When comparing the performance of the final model with the previous attempts, there is a noticeable improvement. Based on Table 2, the MDL model achieved a 54% accuracy, which is an additional 5.26% compared to the SDA using only concatenation of features (48.95%). This time results were statistically significant with p=0.008 at α=0.05. Furthermore, we noticed a reduced variability with regards to how the model performed on the different folds.

4. Discussion

In the first experiment we compared single-layer DAs with single-layer MLPs in order to assess the effects of latent feature discovery. Focusing on the mean accuracy, we noted an increase of 2.63% after using DAs. However, statistical analysis revealed that the improvements were not significant. A possible explanation is that although latent features were discovered, they lacked sufficient discriminative ability due to the model not being able to learn inter-modal features due to the shallow structure. For the next investigation we attempted to use a deeper structure.

The outcome for the SDA experiment was surprising since deep learning is usually expected to lead to improvements, however based on these results it was instead detrimental. Statistical testing shows that this was not significant (p=0.19), although it still meant that there was no observable advantage to simply applying a deeper model. It was apparent that the problem was with how the data fusion was handled. In the succeeding experiment, we corrected the approach by applying MDL.

Based on the results in Table 2, significant performance were made by applying a MDL strategy. The final MDL model achieved a 54% accuracy, an additional 5.26% compared to the SDA, and a statistically significant improvement with p=0.008 at α=0.05. In addition, there was less variability with regards to how the model performed on the

Table 2: Comparison of Performance Results

Model	Fold										Mean	Variance
	1	2	3	4	5	6	7	8	9	10		
MLP	47.37%	47.37%	47.37%	47.37%	47.37%	47.37%	47.37%	52.63%	47.37%	47.37%	**47.89%**	2.77%
DA	47.37%	52.63%	52.63%	47.37%	52.63%	47.37%	47.37%	52.63%	52.63%	52.63%	**50.53%**	7.39%
SDA	47.37%	47.37%	47.37%	47.37%	47.37%	47.37%	57.89%	47.37%	47.37%	52.63%	**48.95%**	12.62%
MDL	57.89%	57.89%	52.63%	52.63%	57.89%	52.63%	52.63%	52.63%	52.63%	52.63%	**54.21%**	6.46%

different folds. These results, indicate that a deep multimodal learning approach is an effective modelling strategy for classifying multimodal stress data.

5. Summary and Conclusion

In summary, this work presented an improved method for building mental stress models using multimodal data and multimodal deep learning. By using a monitoring platform and unobtrusive wearable sensors, data was gathered from subjects engaged in authentic work activities. Psychology-based annotation tools collected stress-related context while wearable sensors tracked physiological signals of heart rate and skin conductance. Then, different structural combinations of autoencoders were tested to discover which could best identify latent features between the physiological and psychological data. Specifically, single-layer denoising autoencoders (DA), a 3-layer stacked denoising autoencoders(SDA), and an SDA with a multimodal deep learning scheme(MDL) were used. All models performed better than the baseline traditional models using standard machine learning methods. The most significant improvements were achieved by applying the MDL. On the other hand, simple feature-level concatenation (i.e., in the SDA) resulted in a slight performance loss compared to single layer DAs. These results support findings from previous works in multimodal learning that state that combinations of certain modalities require separate feature learning phases to discover unimodal features and multimodal features. All models performed better than random despite using naturalistic work activity data without artificially injected stressors. Furthermore, significant performance gains were achieved by applying a multimodal deep learning strategy compared to all other tested approaches. These results show that deep multimodal learning is an effective method of building psycho-physiological stress models.

References

[1] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49:1017–1034, 2013.

[2] C. S. Carver, M. F. Scheier, and J. K. Weintraub. Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology*, 56(2):267–83, 1989.

[3] H. Iso, C. Date, A. Yamamoto, H. Toyoshima, N. Tanabe, S. Kikuchi, T. Kondo, Y. Watanabe, Y. Wada, T. Ishibashi, H. Suzuki, A. Koizumi, Y. Inaba, A. Tamakoshi, and Y. Ohno. Perceived mental stress and mortality from cardiovascular disease among Japanese men and women: the Japan Collaborative Cohort Study for Evaluation of Cancer Risk Sponsored by Monbusho (JACC Study). *Circulation*, 106(10):1229–1236, Sep 2002.

[4] Paolo Melillo, Marcello Bracale, and Leandro Pecchia. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *BioMedical Engineering OnLine*, 10:96+, November 2011.

[5] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, Bellevue, USA, June 2011.

[6] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott Fisk, Fernando De la Torre, Asim Smailagic, Daniel P Siewiorek, Mustafa alAbsi, Emre Ertin, et al. Personalized stress detection from physiological measurements. *International symposium on quality of life technology*, pages 28–29, 2010.

[7] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In *Mobile computing, applications, and services*, pages 211–230. Springer Berlin Heidelberg, 2012.

[8] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December 2010.

[9] W. H. Vogel. Coping, stress, stressors and health consequences. *Neuropsychobiology*, 13(3):129–135, 1985.

[10] J. Zhai and A. Barreto. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, pages 1355–1358, August 2006.