A BAG-OF-WORDS MODEL FOR TASK-LOAD PREDICTION FROM EEG IN COMPLEX ENVIRONMENTS

Lenis Mauricio Meriño[1], Jia Meng[3], Stephen Gordon[4], Brent J. Lance[5], Tony Johnson[4], Victor Paul[6], Kay Robbins[2], Jean M. Vettel[5], Yufei Huang[1]

1. Department of Electrical and Computer Engineering, 2. Department of Computer Science, University of Texas at San Antonio, USA
3. Picower Institute for Learning and Memory, Massachusetts Institute of Technology, USA
4. DCS Corporation, USA, 5. U.S. Army Research Laboratory, USA
6. U.S. Army Tank Automotive Research, Development and Engineering Center, USA

ABSTRACT

Neurotechnologies based on electroencephalography (EEG) and other physiological measures to improve task performance in complex environments will require tools and analysis methods that can account for increased environmental noise and task complexity compared to traditional neuroscience laboratory experiments. We propose a bag-of-words (BoW) model to address the difficulties associated with realistic applications in complex environments. In this paper, our proof-of-concept results show that a BoW classifier can discriminate two task-relevant states (high versus low task-load) while an individual performs a simulated security patrol mission with complex, concurrent tasking. Classifier performance is largely consistent across six simulation missions for a given participant, but performance decreases when trying to predict between two individuals. Overall, these initial results suggest that this BoW approach holds promise for detecting task-relevant states in real-world settings.

Index Terms— Electroencephalography (EEG), Bag-of-words (BoW) model, Participant task-load prediction.

1. INTRODUCTION

Interpreting brain states to improve task performance based on EEG recordings comprises one of the most active areas of research in brain-computer interfaces [1, 2, 3]. Several studies have successfully predicted events in EEG data recorded in well-controlled laboratory settings where participants are confined to perform a single task with events presented at carefully timed intervals and unwanted noise sources, such as eye activity and body movements, are intentionally minimized [4, 5]. Here, we aim to extend this work to classify task-relevant states in less tightly-scripted experiments and a more complex environment. Using data from a simulated security patrol mission, we investigate the prediction of two task-relevant states with concurrent tasking and variable event timing.

In our dataset [6, 7], each participant performed the role of a Vehicle Commander in six simulated low-threat patrol missions while EEG was recorded. In each mission, the Commander was responsible for multiple, concurrent tasks (see Section 2) while he navigated among three checkpoints in the simulated city (Figure 1A). The frequency and difficulty of these tasks varied throughout the mission, providing time periods with infrequent tasks and less visual and auditory information to process (low task-load) and periods with high frequency task occurrence and overlapping visual and auditory tasks to manage (high task-load). Six sections were identified in each mission that corresponded to three sections of expected low task-load (L1-L3) and three sections of expected high task-load (H1-H3) as shown in Figure 1A; however, within each of these sections, the specific experimental events, their timing, and their durations fluctuated based on the inherent dynamics of a patrol mission. In addition, the Commander experienced the simulated vehicle movement on a ride motion simulator and freely moved his eyes, head, and arms to interact with two touchscreen interfaces that controlled the environment (Figure 1B). These design elements helped immerse the participant into the simulated environment [6], but they also introduced large non-brain signals into the dataset that are pervasive across the task [8]. Consequently, the complexity of the experimental tasks and the increased movement artifacts potentially makes the prediction of task-relevant states very challenging.

We propose in this paper to apply the bag-of-words (BoW) model, a method widely used in computer vision, for EEG classification. The use of the BoW model is motivated by the similarity of our problem to those in computer vision, where the number, the spatial locations, and the scale of patterns in two images under the same category can be very different. This mirrors one inherent challenge of brain state classification where the neural responses to task-relevant

events vary within the same task-relevant state. Our analyses examined the performance of the BoW model on two participants. The first analysis looked at the variability of the optimal BoW model for one participant across five missions. The second analysis investigated if the trained classifiers were participant-dependent, while the third analysis examined if the predictive patterns for one participant are consistent across missions. Collectively, our results showed that the proposed BoW model produced good performance in differentiating low and high task-load sections, providing a proof-of-concept of this approach for detecting task-relevant states in real-world settings.

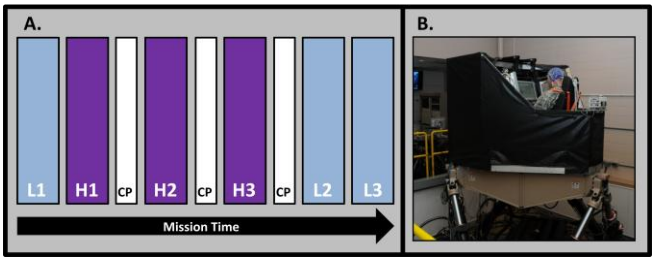

Figure 1(A) Overview of mission: there are three low task-load sections (L1-L3, light blue), three high task-load sections (H1-H3, purple), and three checkpoints (CP, white). (B) Soldier sitting on a six degree-of-freedom Ride Motion Simulator to simulate realistic movements of a vehicle. Photo provided by Detroit Arsenal Media Services

2. EXPERIMENT AND DATA

A total of 14 U.S. Army Sergeants, all male combat veterans of Iraq or Afghanistan, participated in the experiment at the Ground Vehicle Simulation Laboratory (GVSL) at the Detroit Arsenal in Warren, MI. Although the experiment was conducted using a Commander-Driver team [6], this analysis focuses only on the Commander, and only two datasets were used in this proof-of-concept analysis. A previous analysis using 4 seconds epoched data, the bandpower of multiple frequency bands (delta, theta, alpha, low beta and high beta) as features and a SVM algorithm, examined only the auditory communications within the mission, showing that classification between irrelevant audio messages and relevant audio messages requiring the Commander to respond could be performed with 67% prediction accuracy [7].

Each participant completed six, low-threat patrol missions (averaging 22 minutes in duration) through a simulated urban desert terrain while experiencing the realistic movements of the vehicle on a ride motion platform (Figure 1B). In each mission, the Commander was responsible for multiple, concurrent tasks, including route planning and navigation, responding to various auditory communications about mission status and coordination, and maintaining local situational awareness to detect and report visual targets. The frequency and difficulty level of these numerous tasks varied throughout the mission, providing time periods with infrequent tasks and minimal sensory information to process (low task-load) and periods with high-frequency task occurrence and overlapping visual and auditory tasks to manage (high task-load). Three sections of each task-load level were analyzed, each approximately 3 minutes in duration (Figure 1A). EEG was recorded with a 64-channel BioSemi system at a sampling rate of 256 Hz, and four additional electrooculography channels were recorded. All channel data was filtered (FIR from 1 to 50 Hz) to remove frequency domain noise, and we decomposed the EEG signals using Independent Component Analysis (ICA) [9] into 68 independent components. Time-frequency decomposition was subsequently applied to the IC data (using a highly popular algorithm based on Morlet wavelet, which is widely use on EEG) resulting in a tensor of spectral power in 4 dimensions: power, IC, frequency, and time. The power in each of the six mission sections was normalized to the length of the respective section.

3. PROPOSED BOW MODEL FOR TASK-LOAD PREDICTION

A BoW model is a method for treating a classification problem as a dictionary, i.e. an unordered set of words. Figure 2 illustrates the three-step process we used to generate a BoW model: (1) identify a set of discriminate features, (2) construct a dictionary, and (3) train a BoW model and assess prediction of task-relevant states.

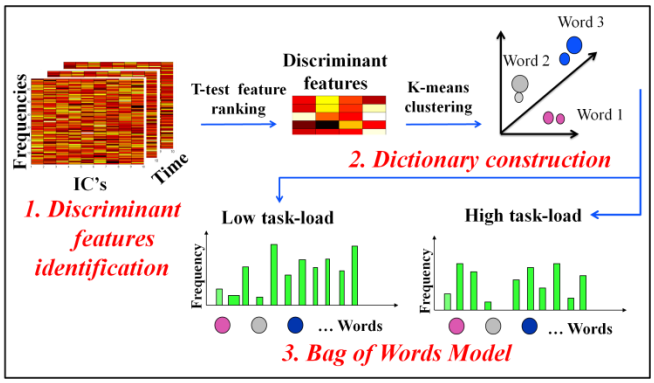

Figure 2 Overview of BoW model generation

3.1. Identification of discriminant features

For the EEG data, a feature is signal power defined at IC, frequency, and time – or $\{c, f, t\}$. In this work, we ignore the temporal dependence of data, and thus a feature vector at t can be defined as $y(t) \in R^{N \times F}$, which contains N (IC) by F (frequencies) normalized powers at time t. Discriminant features are identified separately for each pair of low and high sections (total of nine L-H combinations) for each participant. A filter-wrapper strategy as proposed in [10] was adopted for feature selection, where an initial feature ranking is determined followed by a selection step moving

down the ranked list. The initial ranking is determined by a t-test for each single feature at (c, f) that captures how well the feature discriminates a low task-load section from a high task-load section (power from time-frequency decomposition). The features are ranked according to ascending order of the absolute value of their t-values. Features are then selected using a sequential forward search that is performed over ranked features, where at each search step, one feature is added to the model, and then prediction performance based on cross validation is assessed after execution of the steps in 3.2 and 3.3 to generate a BoW model for that set of features. 10-fold cross validation was performed on the EEG data by dividing it into epochs of 500 samples (approximately 2 seconds of data). This three step process was done parametrically for dictionary sizes of 2-10 words and from 1-300 total features.

Thus, the selected feature set for a given L-H discrimination are the top ranked features that achieve the highest performance prediction (lowest mean error rate, red dot in Figure 3). The BoW model based on K clusters of discriminant features is the final model.

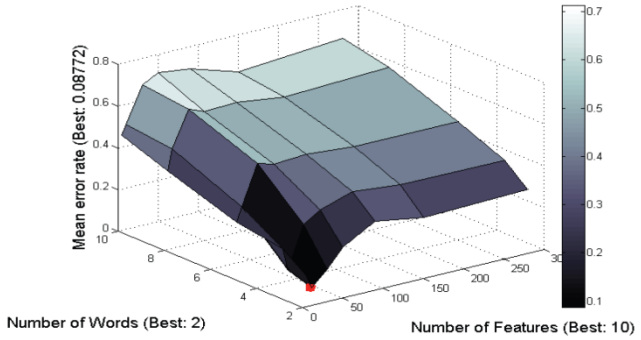

Figure 3: Plot of the error rate as a function of the numbers of words and top ranked features for L2 vs. H3 for participant 1. For this L-H pair, the best error rate is achieved at 2 words and 10 top features.

3.2. Construction of dictionary

Given a set of discriminant features, a dictionary D consisting of an unordered set of "words," or significant patterns of the feature vector, is created from the training data that is independent of L and H class labels. The patterns are defined as the most representative values of the feature vector. To identify the patterns, a k-means clustering is applied, and the patterns are taken as the centroid of the clusters. The optimal number of the clusters, K, was determined by cross validation as described in Section 3.1.

3.3. Training and Task-load classification

When training a BoW model, the goal is to calculate the respective distributions of words in the high/low task-load mission sections. The distribution of words for a section is calculated by mapping each data sample y_t in the section to the nearest word in D via Euclidian distance. The distribution is taken as the count of the mapped words in the section. Performance prediction is assessed using samples of EEG data with length T. For a sample $y_t, \forall t$ is first mapped to the respective nearest word pattern in D. Let $x_t \in \{1, \dots, K\}$ represent the word index for y_t. Prediction is then carried out with a Naive Bayes classifier, where the posterior probability of task-load type i (L or H) given the data, or $x_{1:T}$, is calculated as

$$p(i|x_{1:T}) \propto \prod_{t=1}^{T} p(x_t|i)\, p(i)$$

$$= \prod_{j=1}^{K} p_{j,i}^{N_j} p(i) \qquad (1)$$

where $p(i)$ is the prior probability of observing task-load type i, N_j is the number of samples whose $x_t = j\ \forall t$, and $p_{i,j}$ is the probability of observing word j in task-load i (obtained from the training). The classifier predicts the task-load to be the type (L or H) that achieves the highest posterior probability $p(i|x_{1:T})\ \forall i$.

4. RESULTS

Results are described for three preliminary analyses of the BoW model. The first analysis looked at the stability of the dictionary and feature set across each pair of low and high sections for participant 1 across five missions. The second analysis examined how well the classifier trained on participant 1 performed on participant 2, and the third analysis examined how well a classifier performed across missions for participant 1.

4.1. Investigating the best set of features and *K*

In the first analysis, we examined the variability of the optimal BoW model across each L-H combination for each mission for participant 1. To this end, the procedure described in 3.1 and shown in Figure 3 was performed for each L-H pair across the participant's five missions (one mission had missing data). Cross validation was performed on each mission independently, but the results summarized in Table 1 are summarized across mission. Across the nine L-H pairs, using 2 words consistently achieved the best error rate. This result implies that there exist two types of feature patterns that define the differences in activity between sections of low and high task-load, although the two particular patterns are not necessarily the same across the L-H pairs. The total number of discriminating features, however, varies substantially between the L-H pairs, with feature numbers ranging from 4 to 280.

4.2. Between-participant performance

In the second analysis, we investigated if the trained classifiers were participant-dependent. To this end, the trained classifiers described in section 4.1 for participant 1

were used to predict the task-load for participant 2. Tests were carried out for the 3 L-H combinations involving L1, and Table 2 summarizes the results for the subset of tested missions. The prediction errors on participant 2 increase considerably for most of the missions, and these initial results suggest that a participant's patterns to task-relevant events vary among individuals. Consequently, classifiers for use in complex environments may need to be trained separately for different participants.

Table 1. Best feature set and word number

Types	Feature #	Word #	Mean error rate
L1 vs. H1	50	2	0.0633
L1 vs. H2	30	2	0.1097
L1 vs. H3	280	2	0.0619
L2 vs. H1	10	2	0.0877
L2 vs. H2	4	2	0.1345
L2 vs. H3	10	2	0.1435
L3 vs. H1	150	2	0.0713
L3 vs. H2	150	2	0.0748
L3 vs. H3	150	2	0.0887

Table 2. Between-participant performance

Task-load, Mission	P#1 error rate	P#2 error rate
L1 vs. H1, 2	0.121	0.145
L1 vs. H1, 3	0.035	0.331
L1 vs. H1, 4	0.041	0.315
L1 vs. H1, 5	0.089	0.227
L1 vs. H1, 6	0.029	0.183
L1 vs. H2, 2	0.141	0.195
L1 vs. H2, 3	0.131	0.235
L1 vs. H2, 4	0.031	0.215
L1 vs. H2, 6	0.094	0.050
L1 vs. H3, 2	0.094	0.125
L1 vs. H3, 3	0.100	0.243
Mean	**0.082**	**0.206**

Table 3. Within-participant mean error rate across missions

Types	Mission 3	Mission 4	Mission 5	Mission 6
L1 vs. H1	0.039	0.006	0.029	0
L3 vs. H3	0.070	0.08	0.379	0.219
L1 vs. H3	0.041	0.006	0.029	0.032

4.3. Cross-mission performance

Finally, we investigated if the predictive patterns to task-relevant events were consistent across the missions. To this end, classifiers were trained for participant 1 using data in mission 2 and then were used to predict task-load in the remaining missions. Table 3 shows the results for three L-H combinations. Overall, the error rates remain small and comparable to those for mission 2 (except mission 5 and 6 of L3 vs. H3). Taken together, we conclude that participant 1 engaged in fairly consistent activities across missions, but future work will need to examine additional mission segments and experimental participants to better assess the within-participant stability of the classifier.

5. CONCLUSION

We used EEG data recorded from a Soldier performing the role of a Commander in a simulated security patrol mission on a ride motion platform to examine whether a bag-of-words (BoW) model could successfully classify task-relevant states despite increased complexity of the experimental tasks and increased noise from movement artifacts and other non-brain signals. Our preliminary results suggest that low and high task-load states may be reliably discriminating with a few stable words, and these patterns are relatively stable across similar missions for the same participant. In the two participants studied, however, the classifier from participant 1 did not perform as well on participant 2, suggesting that the task-relevant patterns may vary among individuals. Future work will examine the contribution of brain and non-brain sources to these predictive, task-relevant patterns.

More generally, our results indicate the promise of the BoW approach for tackling one of the inherent challenges of brain state classification when the number, spatial locations, and scale of neural responses to task-relevant events vary within the same task-relevant brain state. That is, BoW does not require tightly-scripted experiments and precise event timing. In fact, BoW is not limited to discrimination of two general brain states. Future work will examine if the BoW model can identify the particular types of events that contribute to the participant's task-load, such as hearing an audio communication, seeing a visual target, or pressing buttons to manipulate one of his touchscreen interfaces. If successful, the BoW approach will provide a critical analysis method needed to move neurotechnology development into more complex environments, providing an avenue to improve task performance in real-world settings.

6. ACKNOWLEDGEMENTS

The authors would like to thank Marcel Cannon, Chris Manteuffel, Matthew Jaswa, Jennifer Ammori, Kelvin Oie, Kaleb McDowell, and our colleagues. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implies, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for the Government purposes notwithstanding any copyright notation herein. This work also received computational support from Computational System Biology Core, funded by the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

7. REFERENCES

[1] F. Lotte, M. Congedo, A. L´ecuyer, F. Lamarche, B. Arnaldi, et al., "A review of classification algorithms for EEG-based brain–computer interfaces," Journal of neural engineering, vol. 4, 2007.

[2] Lance, B.J.; Kerick, S.E.; Ries, A.J.; Oie, K.S.; McDowell, K.; , "Brain–Computer Interface Technologies in the Coming Decades," Proceedings of the IEEE , vol.100, no.Special Centennial Issue, pp.1585-1599, May 13 2012.

[3] Zander, T.O., Kothe, C. "Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general," J. Neural Eng., Vol. 8, No. 2, 2011.

[4] J. Meng, L.M. Meriño, N.B. Shamlo, S. Makeig, K. Robbins, and Y. Huang, "Characterization and robust classification of EEG signal from image rsvp events with independent time-frequency features," PLOS ONE, vol. 7, no. 9, pp. e44464, 2012.

[5] Allison, B.Z., Polich, J. "Workload assessment of computer gaming using a single-stimulus event-related potential paradigm," Biological psychology, Vol. 77, No. 3, pp. 277-283, 2008.

[6] Vettel, J., Lance, B., Manteuffel, C., Jaswa, M., Cannon, M., Johnson, T., Paul, V., & Oie, K. (2011) Mission-Based Scenario Research: Experimental Design and Analysis. Proceedings of Modeling and Simulation, Testing and Validation (MSTV) MiniSymposium. At the Ground Vehicle Systems Engineering and Technology Symposium (GVSETS) 2011. Reprinted as ARL-RP-0352.

[7] B. Lance, S. Gordon, J. Vettel, T. Johnson, V. Paul, C. Manteuffel, M. Jaswa, and K. Oie, "Classifying highnoise EEG in complex environments for brain-computer interaction technologies," Affective Computing and Intelligent Interaction, pp. 467–476, 2011. Reprinted as ARL-RP-0350.

[8] Kerick, S.E., Oie, K.S., & McDowell, K. (2009) Assessment of EEG Signal Quality in Motion Environments. ARL-TR-4866.

[9] Bell, A.J., Sejnowski, T.J., "An Information-Maximization Approach to Blind Separation and Blind Deconvolution", Neural Computation, vol. 7, pp. 1129-1159, 1995

[10] R. Ruiz, J.C. Riquelme, and J.S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," Pattern Recognition, vol. 39, no. 12, pp. 2383–2392, 2006.