Single-trial identification of failed memory retrieval

Eunho Noh[1], Matthew V. Mollison[2], Tim Curran[3], and Virginia R. de Sa[4]

Abstract— **We show that it is possible to distinguish unsuccessful from successful retrieval of study items based on single-trial EEG recorded during the test phase of 3 separate recognition memory experiments. The overall classification accuracy across all 34 classification problems was 58.4 %. The classification accuracy monotonically increased to 68.03 % by only classifying trials with high classifier confidence levels. The likelihood of remembering a study item for trials with the 10% highest and lowest classifier outputs were 0.80 and 0.45 respectively. This suggests that the classifier outputs are reflecting the level of retrieval during the test phase. These findings combined with previous single-trial results predicting subsequent memory from EEG recorded during and prior to memory encoding will provide a basis for a passive brain-computer interface (BCI) system for improving memory.**

I. Introduction

Brain computer interfaces (BCIs) are devices that allow interaction between humans and computers using the brain signals of the user. There are three types of BCI system based on what kind of control signal the system utilizes. In active BCI systems, the system outputs are obtained by interpreting the brain activity which is directly and consciously controlled by the user. In reactive BCI systems, the system outputs are obtained by interpreting the brain activity which results from a reaction to external stimulation. The brain activity is indirectly modulated by the user with the objective of intentionally controlling an application. In a passive BCI system, outputs are derived by interpreting brain activity where there is no voluntary control, to enrich the user's interaction with the system.

In [9], we proposed a passive BCI system on deciphering neural correlates of memory from single-trial EEG. This system had three major components: 1) encoding preparedness: the system monitors the brain activity of a user and predicts the user's preparedness for learning to present study items at estimated *optimal* times, 2) encoding success: the system utilizes the brain activity during encoding to predict the success of the encoding process, and 3) confidence at retrieval: the system extracts information related to the level of reinstatement from the brain activity during representation (or test) of an item.

Single-trial classification results for the first two aspects of the system has been previously published in [7]. In [7], pattern classifiers were trained to learn the temporal and spectral differences in the brain activity during memory encoding between the subsequently remembered vs. forgotten trials. By combining the pre- and during-stimulus predictions, Noh and colleagues were able to predict whether a given study item would be remembered in the test phase with 59.64 % accuracy. In [8], the temporal information in EEG activity during memory retrieval was used to predict correctly identified old vs. correctly rejected new trials. The authors were able to distinguish whether the subject identified the studied or foil items correctly with an average accuracy of 61 % using data from 3 separate recognition memory experiments. It was found that the classifier outputs (or classifier scores) reflected the amount of information retrieved from the study episode. In this paper, we use data and classification methods used in [8] to predict unsuccessful from successful retrieval of study items based on single-trial EEG during memory retrieval. This classification problem corresponds to the third and final component of the passive BCI system proposed in [9].

II. The dataset

EEG for the analysis was previously recorded in 3 separate visual memory task experiments [6]. In a typical recognition memory experiment subjects are given a list of study items in the study phase. In the test phase, the studied items are given with unstudied foil items and the subjects are instructed to distinguish the studied items from the foil items using some sort of response. There are four basic behavioral categories corresponding to the test trials for a recognition memory experiment.

1) Hit: correctly identified old (studied) items.
2) Correct rejection: correctly rejected new (foil) items.
3) False alarm: new (foil) items with old responses.
4) Miss: old (studied) items with new responses.

A. Experimental paradigm

In the current experiments, the study items (color images of physical objects, animals, and people) were given with extrinsic source information. In addition to memorizing the given study items, the subjects were told to associate the specific extrinsic source given with each study item. Two types of sources were considered for the experiments. In experiment 1, study items were given on either the left or right side of the computer screen. In experiment 2, the study items were given in color frames out of 8 possible colors. In experiment 3, the two source types were considered separately in different blocks. Experiment 3 was conducted in two separate sessions on different days where the two

[1]Eunho Noh is with the Department of Electrical & Computer Engineering, UCSD eunoh@ucsd.edu

[2]Matthew V. Mollison is with the Department of Psychology and Neuroscience, University of Colorado Boulder

[3]Tim Curran is with Faculty of Psychology and Neuroscience, University of Colorado Boulder

[4]Virginia R. de Sa is with Faculty of Cognitive Science, UCSD

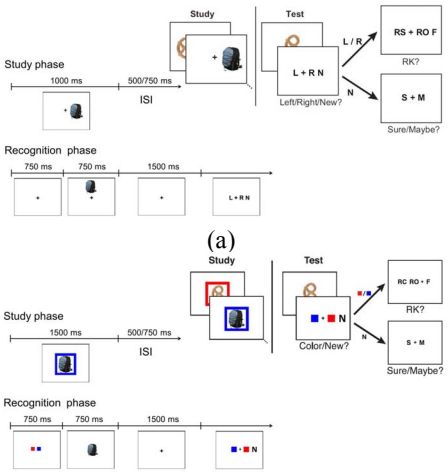

Fig. 1: The experimental paradigm for (a) the location source experiments and (b) the color source experiments. For each figure, the timings of the study phase and recognition (or test) phase are given on the left and an illustration of the study and test tasks are given on the right hand side.

source conditions were given on both days. An illustration of the experimental paradigm is given in Figure 1.

In the study phase, the subjects learned the pictures and their associated source information. In the test phase, they had to distinguish the studied pictures from random distractors with the appropriate location or color using two consecutive responses. The subjects gave an source 1/source 2/new decision in the first response and a subjective rating of the decision in the second response. The following breaks down the test phase procedure for a given test item:

- The item is recognized as old.
  - First response
    The source information corresponding to the item is selected: L/R (left/right) for location source; color of the frame for color source.
  - Second response: A subjective rating on the source judgment is given with one of the following options.
    - Remember source (RS): The subject believes he/she correctly remembers the source information.
    - Remember other (RO): The subject remembers something other than the given source information.
    - Familiar (F): The item looks familiar.
- The item is recognized as new.
  - First response: A new (N) response is given.
  - Second response: A subjective rating on the new judgment is given with either maybe (M) or sure (S).

Based on the correctness of the source/item judgments and subjective ratings, the test trials can be divided into 13

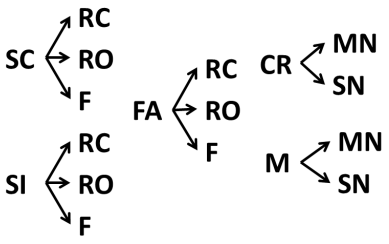

Fig. 2: Categorization of the trials based on the subjects' source judgments (SC: source correct, SI: source incorrect, CR: correct rejection) and subjective ratings (RS: remember source, RO: remember other, F: familiar, MN: maybe new, SN: sure new).

conditions as illustrated in Figure 2. More details on the experimental paradigm can be found in [6]. In this paper, we were only interested in the correct source retrieval (SC) and miss (M) trials.

B. EEG acquisition and pre-processing

EEG was recorded with a Geodesic Sensor Net™ (HydroCel GSN 200, v. 2.1; [12]) (250 Hz sampling rate for Experiments 1 and 2, 500 Hz sampling rate for Experiment 3 and subsequently downsampled to 250 Hz) using an AC-coupled 128-channel, high-input impedance amplifier (300 MΩ, Net Amps™; Electrical Geodesics Inc., Eugene, OR). Initial common reference was the vertex channel (Cz) and the individual electorates were adjusted until impedance measurements were lower than 40 kΩ. EEG epochs from the recognition phase of each experiment were extracted and recalculated to average reference. Each epoch was filtered between 0.1 and 50 Hz using a 40 tap FIR filter and baseline corrected using data from -200-0 ms.

III. Classification

There are two well known ERP (event-related potential) effects related to memory retrieval where correctly identified old trials show significantly different ERPs from the correctly rejected new trials. This difference in ERP is commonly referred to as the old/new effect. The parietal old/new effect is a positive going ERP observed between 500-800 ms in the parietal electrodes. It shows greater amplitude for the correctly recognized old compared to the correctly rejected new items [15], [14], [3], [11]. It has been found that this effect correlates with the amount of information retrieved from the study episode [13]. The frontal old/new effect is a negative-going ERP observed between 300-500 ms in the frontal electrodes (hence often referred to as the FN400). This ERP typically shows a more negative peak for less familiar items while it shows no difference for different amounts of episodic information.

For the current analysis only the test trials given in the study list were considered (in other words, the trials where foil items were given were not considered). The classifiers were trained to find the projection function onto the vector

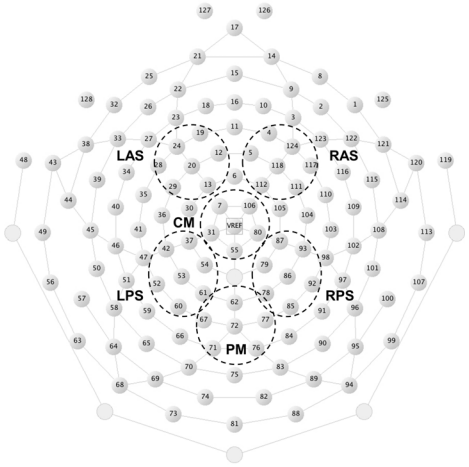

Fig. 3: The GSN electrode locations used to record the EEG and the six channel groups on which classification analysis was conducted. LAS *left anterior superior*, RAS right anterior superior, CM *central medial*, LPS *left posterior superior*, RPS right posterior superior, and PM *posterior medial*.

perpendicular to the decision boundary which represents the amount of information retrieved from the study episode. A two-class binary classifier with probability outputs ($0 \leq p \leq 1$) was trained to discriminate between correctly identified old trials with correct source retrieval (*class 1: SC-RS, SC-RO*) and incorrectly rejected old trials (*class 2: M*) for each subject (as illustrated in Figure 2). In order to maximize the difference in retrieved information between the two classes, the SC-F (correct source retrieval with familiar judgments) trials were not included in class 1 since they were likely to include many guesses on the source decision. The individual classifier performances were evaluated on class 1 vs. class 2 classification using a balanced cross-validation procedure[1].

As done in [8], the features for the classifier were selected based on previous findings on the old/new effect. The EEG data between 300 and 800 ms from six channel groups were selected (LAS, RAS, CM, LPS, RPS, and PM as given in Figure 3) in order to include the electrode locations and time segments where the two old/new effect would be expected. The 30-dimensional feature vector for a given trial was computed from the selected data by averaging over 100 ms length non-overlapping windows (5 time windows $\times$ 6 channel groups = 30 features). A LDA (linear discriminant analysis) classifier was trained to distinguish these feature vectors. The classifier was calibrated to give probability outputs (or classifier scores) based on a permutation test [4]. These values correspond to the likelihood of a given trial being retrieved with the correct source information. For initial classification, a given trial was classified as correctly retrieved if the classifier score was larger than 0.5 and classified as incorrectly rejected otherwise.

Classifier training and evaluation were conducted separately on each subject. Data from Experiment 3 were divided into the location source blocks (denoted as Exp 3-loc), and color source blocks (denoted as Exp 3-col) based on results which found different characteristics in the brain activity related to memory retrieval between location and color source retrieval [8]. Subjects who had less than 50 trials within each class after artifact trial rejection were not included in the classification analysis. This resulted in 34 individual classification problems (11 from Experiment 1, 2 from Experiment 2, 12 from Exp 3-loc, and 9 from Exp 3-col). Note that Exp 3-loc and Exp 3-col datasets were multi-session datasets as described in Section II-A.

In order to verify whether the classification accuracy was significantly over chance (50 %), the 95 % confidence interval for chance level performance was computed *for each subject* based on the number of trials in classes 1 and 2. This was done by using Wald intervals with small sample size adjustments [1]. The results were considered to be significantly over chance only when the accuracy was over this threshold.

IV. Results

A. Classification accuracy

The overall classification accuracy across the 34 classification problems (calculated for all trials from all the available subjects) was 58.40 % and the individual accuracies were significantly over chance (significantly over 50 % with $p < 0.05$) for 20 out of the 34 results with none going significantly below 50 %. The overall accuracy for the single session datasets and multi-session datasets were 58.92 % and 57.99 % respectively. These values were not significantly different based on a one-sided rank sum test. The individual classification results are given in Figure 4 (a).

B. Classification accuracy based on the classifier scores

Additional analysis was conducted to investigate whether the classification accuracy increased by ignoring the trials with ambiguous classifier scores (trials with classifier scores close to 0.5). This was done by sorting the trials by classifier score given by the retrieved vs. miss classifier and only selecting the top/bottom trials for classification. Note that the classifier scores were not recomputed for this analysis. The trials with classifier scores close to 0.5 were excluded from classification. Selection ratios of 10, 20, 30, 40, and 50 % were chosen (a selection ratio of 50 % means all available trials were considered for classification and a selection ratio of 40 % means the top and bottom 40% were considered, etc). The overall accuracy increased monotonically as less of the ambiguous trials were considered for classification (see Figure 5). The overall accuracy for the top/bottom 10 % of the trials was 68.03 % (68.9 % for the single session datasets and 67.3 % for the multi-session datasets). The individual classification results when the top/bottom 10 % of the trials were considered for classification are given in Figure 4 (b).

[1]In a balanced leave-two-out cross-validation procedure, two trials (one from each class) are randomly selected as the validation set and left out of any training procedure and used as the validation set. This procedure is done for each fold until all trials are used for validation. Classifier performance is evaluated by averaging over the accuracies across all folds.

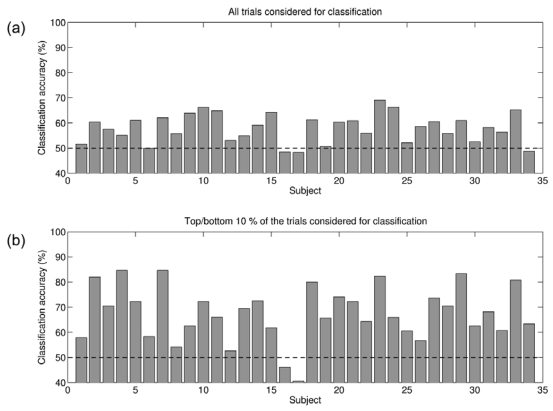

Fig. 4: The individual classification accuracies (a) when all trials are considered for classification and (b) the 10 % of the trials with the highest/lowest classifier scores are considered for classification.

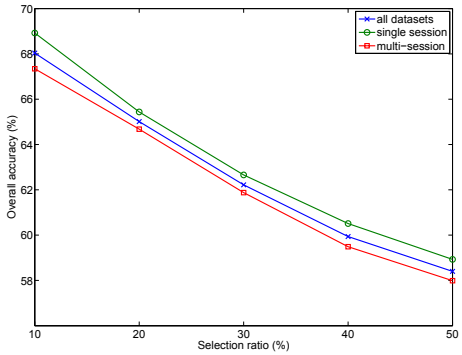

Fig. 5: The change in classification accuracy as less trials with classifier scores close to 0.5 are considered for classification.

The likelihood of correctly retrieving a study item for trials with the 10 % highest and lowest classifier scores were 0.8 and 0.45 respectively.

C. Visualization of the ERP components

The ERP components for the two classes were computed to visualize the information utilized by the classifier. Due to the space limit, we only give the ERPs from the two multi-session datasets (Exp 3-loc and Exp 3-col). However, the single-session ERPs showed similar effects with a larger difference between the two classes. The amplitude difference between the two classes were evident in both the frontal (LAS and RAS) and posterior channel groups (LPS and RPS). However the average amplitude difference in the two frontal channel groups was significantly larger than the two posterior channel groups between 300-800 ms ($p < 0.005$).

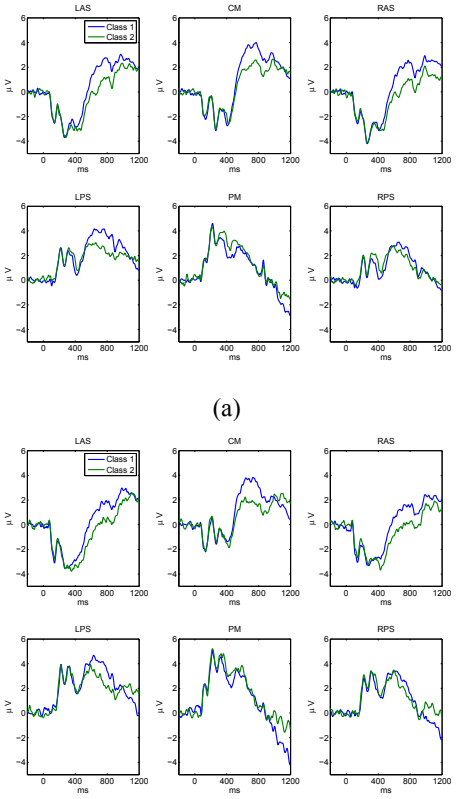

Fig. 6: The ERPs waveforms from the 6 channel groups used for classification. (a): Exp 3-loc; (b): Exp 3-col (class 1: source correct (SC) trials with remember source (RS) or remember other (RO) judgments; class 2: Misses (M)).

V. Discussion

These results show that it is possible to successfully predict successful vs. failed memory retrieval from single-trial scalp EEG activity recorded during memory retrieval with 58.4 % accuracy. Classification was fairly successful even when the dataset was collected from multiple sessions. The prediction rate improved to 68.03 % (a 16 % improvement), by only considering the top/bottom 10 % of the trials for classification. Single-trial classification of episodic retrieval has been investigated using fMRI (functional MRI) [10]. Multi-voxel pattern analysis (MVPA) on 80 distinct anatomical ROIs (regions of interest) revealed that it is possible to discriminate between hits vs. misses with approximately 70 % accuracy. The classification accuracy increased monotonically to 90 % when the top 10 % most confidently classified trials were used to evaluate the classifier. Although the classification accuracy of the EEG-based classifier is lower than the fMRI results, the monotonic increase in classification accuracy (as illustrated in Figure 5) shows that the confidence of the classifier is representative of how well a given trial would be successfully retrieved.

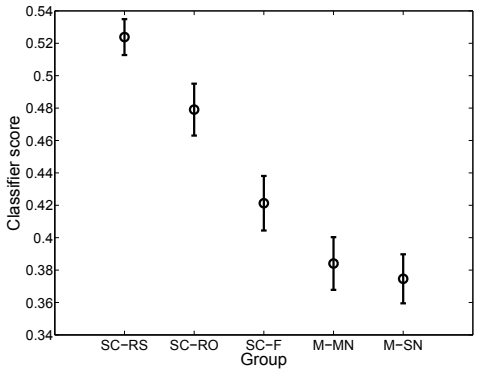

Fig. 7: The estimated means and the approximate 95 % confidence intervals of the classifier scores [5] for the 5 different behavioral conditions.

In order to investigate this matter further, we compared the classifier scores across the 5 different subjective rating conditions given to the trials in classes 1 and 2. The behavioral conditions for the trials in class 1 were remember source (RS), remember other (RO), and familiar (F). We denote these conditions as SC-RS, SC-RO, and SC-F. Note that the SC-F trials were not included in class 1. The behavioral conditions for the trials in class 2 were maybe new (MN) and sure new (SN). We denote these conditions as M-MN, M-SN. All conditions showed significantly different means based on a repeated measure ANOVA ($p < 0.003$) except for the comparison between Miss-M and Miss-S ($p = 0.72$). The classifier scores followed the subjects' subjective ratings of the amount of information they were retrieving from the study episodes as illustrated in Figure 7 suggesting that the classifier scores reflect the subjects' subjective rating on their memory retrieval performance.

Based on these findings we may be able to develop a classifier which can extract information related to the user's confidence during re-presentation of a study item to assess the level of reinstatement without requiring any explicit input (on their confidence) from the user. The items with insufficient reinstatement can be given again for a third time to ensure encoding. to ensure encoding [2]. This classifier can be combined with other classifiers which predict optimal brain states for memory encoding and successful/unsuccessful memory encoding as a passive BCI system for assisting memory formation and retention of the user.

ACKNOWLEDGMENT

This research was funded by NSF grant # CBET-0756828, NIH Grant MH64812, NSF grants # SBE-0542013 and # SMA-1041755 to the Temporal Dynamics of Learning Center (an NSF Science of Learning Center), and the KIBM (Kavli Institute of Brain and Mind) Innovative Research Grant.

REFERENCES

[1] A. Agresti and B. Caffo, "Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures," *The American Statistician*, vol. 54, no. 4, pp. 280–288, 2000.

[2] N. J. Cepeda, N. Coburn, D. Rohrer, J. T. Wixted, M. C. Mozer, and H. Pashler, "Optimizing distributed practice: theoretical analysis and practical implications," *Exp Psychol*, vol. 56, pp. 236–246, 2009.

[3] T. Curran, "Brain potentials of recollection and familiarity," *Memory & Cognition*, vol. 28, pp. 923–938, 2000.

[4] L. Dümbgen, B.-W. Igl, and A. Munk, "P-values for classification," *Electronic Journal of Statistics*, vol. 2, pp. 468–493, 2008.

[5] Y. Hochberg and A. C. Tamhane, *Multiple comparison procedures*. New York, NY, USA: John Wiley & Sons, Inc., 1987.

[6] M. V. Mollison and T. Curran, "Familiarity in source memory," *Neuropsychologia*, vol. 50, pp. 2546–2565, 2012.

[7] E. Noh, G. Herzmann, T. Curran, and V. R. de Sa, "Using single-trial eeg to predict and analyze subsequent memory," *NeuroImage*, vol. 84, pp. 712–723, 2014.

[8] E. Noh, M. V. Mollison, T. Curran, and V. R. de Sa, "Using single-trial analysis to predict and analyze source memory retrieval," submitted to Neuroimage.

[9] E. Noh, M. V. Mollison, G. Herzmann, T. Curran, and V. R. de Sa, "Towards a passive brain computer interface for improving memory," 2014, submitted to the 6th International Brain-Computer Interface Conference 2014.

[10] J. Rissman, H. T. Greely, and A. D. Wagner, "Detecting individual memories through the neural decoding of memory states and past experience," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 9849–9854, 2010.

[11] M. D. Rugg and T. Curran, "Event-related potentials and recognition memory," *Trends in Cognitive Sciences*, vol. 11, pp. 251–257, 2007.

[12] D. M. Tucker, "Spatial sampling of head electrical fields: The geodesic sensor net," *Electroencephalography and Clinical Neurophysiology*, vol. 87, pp. 154–163, 1993.

[13] K. Vilberg, R. Moosavi, and M. Rugg, "The relationship between electrophysiological correlates of recollection and amount of information retrieved," *Brain Research*, vol. 1122, pp. 161–170, 2006.

[14] E. L. Wilding, "In what way does the parietal erp old/new effect index recollection?" *International Journal of Psychophysiology*, vol. 35, pp. 81–87, 2000.

[15] E. L. Wilding and M. D. Rugg, "An event-related potential study of recognition memory with and without retrieval of source," *International Journal of Psychophysiology*, vol. 119, pp. 889–905, 1996.