Combining Computer and Human Vision into a BCI: Can the whole be greater than the sum of its parts?

Eric A. Pohlmeyer, David C. Jangraw, Jun Wang,
Shih-Fu Chang, *Fellow, IEEE* and Paul Sajda, *Senior Member, IEEE,*

Abstract— **Our group has been investigating the development of BCI systems for improving information delivery to a user, specifically systems for triaging image content based on what captures a user's attention. One of the systems we have developed uses single-trial EEG scores as noisy labels for a computer vision image retrieval system. In this paper we investigate how the noisy nature of the EEG-derived labels affects the resulting accuracy of the computer vision system. Specifically, we consider how the precision of the EEG scores affects the resulting precision of images retrieved by a graph-based transductive learning model designed to propagate image class labels based on image feature similarity and sparse labels.**

I. INTRODUCTION

The enormous growth in both computer processing and storage capabilities mean that we now have more information available at our fingertips than ever before. However, the speed at which this information can be accessed far exceeds a human's natural ability to process it, making new techniques that help people sort huge amounts of data and locate only the most relevant extremely important. For example, services such as Flickr and Google Image provide access to billions of images, but locating just a specific type of image among such vast resources is a significant challenge.

The somewhat complementary strengths and weaknesses of human vision and computer vision (CV) offer an excellent opportunity to develop Brain-Computer Interfaces (BCIs) that integrate the two for image retrieval. BCI research has already made significant progress using noninvasive electroencephalogram (EEG) recordings to control devices such as communication systems, computer cursors, and muscle stimulators [1], [2], [3], [4]. This research has both reduced the cost and increased the practicality of noninvasive BCI systems. However, EEG recordings are inherently noisy, and the low signal to noise currently attainable for EEG-BCI control signals typically restricts their use to providing assistive technology to people with disabilities. In a BCI that combines CV with human vision, the CV could be used to refine the output of an EEG-based image detector to both reduce noise and determine the general image characteristics of interest. The CV component could then quickly analyze additional images, returning only the most relevant to the human vision component for evaluation. Fig. 1 shows a BCI system architecture where a user can quickly look at a set of images (flashed at perhaps 5-10 images per second) randomly drawn from a large image database, with images that captured their attention evoking an EEG signature. This signature could be used to label an example set of interesting images that is given to an image retrieval system, which would then reorganize the full database (of conceivably millions to billions of images) so that novel images that were visually similar to those evoking the EEG signature are at the top. This would provide a general purpose BCI system capable of both determining what a user was interested in (within any given image database), and then expediting a search of the database to find additional relevant images.

In this paper we consider the specific issue of interfacing an EEG decoding system with a computer vision system using the hybrid approach described above. Specifically, we investigate the output precision performance of an EEG image detector relative to the input precision requirements of a graph-based transductive learning model designed to detect relevant images based on image feature similarity and sparse labels.

II. METHODOLOGY

A. Image Database

The imagery used to test the EEG and CV systems was taken from the Caltech-101 database [5]. To prevent fluctuations in image size from influencing subjects' visual responses, only a subset of Caltech-101 images were used. Specifically, 62 image categories (a category being images of a common type, e.g. 'elephants', 'grand pianos') were

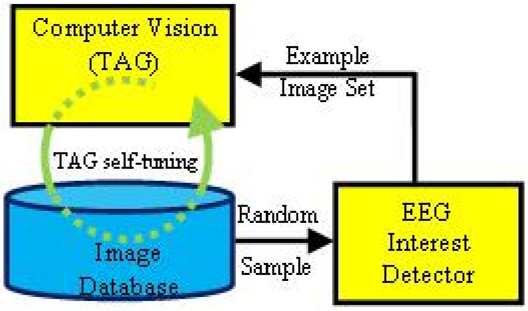

Fig. 1. BCI architecture combining EEG and computer vision components.

Manuscript received April 23, 2010. This work was supported by DARPA under Contracts N10PC20050 and NBCHC080029.

E. A. Pohlmeyer, D. C. Jangraw and P. Sajda are with the Department of Biomedical Engineering, Columbia University, New York, NY 10027 USA ep2473@columbia.edu, dcj2110@columbia.edu, psajda@columbia.edu

J. Wang and S.-F. Chang are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA jwang@ee.columbia.edu, sfchang@ee.columbia.edu

selected on the basis of their images being similar enough in size to be rescaled to uniform dimensions during the visual presentation with negligible distortion. This provided 3798 images (42% of the total Caltech-101 imagery), with, on average, 61 images per category (STD=22, range=31-128 images per category) for the testing database.

B. Human Vision: Object Detection using EEG during Rapid Visual Serial Processing

To identify interesting images, the subjects' EEG activity was mapped onto an 'interest' score. EEG data were recorded using a 64-electrode Biosemi (BrainProducts, Germany) system in a standard 10-20 electrode montage. Data were collected at 2048 Hz, with 60 Hz notch and 0.5 Hz high pass filters. Images were presented in a Rapid Serial Visual Presentation (RSVP) paradigm [6] in which 100 images are displayed at 5 Hz, with self-paced rest periods (typically a few seconds) between the 100-image blocks.

The EEG interest detector has been previously described [7], [8]. Briefly, it is based on a linear model,

$$y_t = \sum_i w_i x_{it} \tag{1}$$

x_{it} is EEG activity at electrode i at time t, and w is a set of spatial (i.e. electrode) weights. Weights are learned from a set of training data so that y maximally discriminates between target and non-target images. Data are binned into windows with a temporal resolution (T) of 100*ms*, and weight vectors, w_{it}, are found for several windows following each image presentation (k is the time window index):

$$y_{kt} = \sum_i w_{ki} x_{it}\,, \quad t = T, 2T, ...(k-1)T, kT \tag{2}$$

The y_{kt} obtained for each time window are combined in a weighted average to provide the final interest score (y_{IS}):

$$y_{IS} = \sum_t \sum_k v_k y_{tk}. \tag{3}$$

The EEG interest detector used 1200*ms* of data following each image (although the first 100*ms* was not included in the detector), with the P300 'oddball' response providing much of the discrimatory information [8], [9]. Fisher Linear Discriminant analysis was used to create the spatial coefficients, w_{ik}, and logistic regression was then used to determine the temporal coefficients, v_k [9].

Data were collected from 8 subjects over 16 experimental sessions (each subject participated in 1-3 sessions). In each session, an interest detector was created from a set of training data and then used for several testing RSVP sequences. The training data consisted of 20-35 image blocks (depending on subject performance) with two target images per block, with training target images being baseball gloves. All the training images (targets and distracters) were taken from the Caltech-256 database [10] so no testing images were seen during training. Each testing RSVP sequence involved five 100-image blocks (images taken radomly from the testing database) in which the subject attended to a specific image category (the target) selected by either the subject or the experimenter. A total of 50 testing RSVPs were collected between all the subjects, with 13 different image categories being used as targets. Some subjects were given additional testing RSVP sequences in which the images shown were determined from their previous results. Informed consent was obtained from all participants in accordance with the Columbia University Institutional Review Board.

C. Computer Vision: Transductive Annotation by Graph

The computer vision (CV) system tested is a method termed Transductive Annotation by Graph (TAG), a semi-supervised learning technique that uses an example set of images for pattern discovery, and then scores all the images in a larger database by propagating the results through a graph structure. Details of the algorithm can be found in [11], but, briefly, it uses an affinity graph to quantify the pairwise similarity between images in a database using a predefined feature space. Then, given a set of example images, $\mathcal{X}$, with noisy labels, $\mathbf{E}$ (noisy meaning the example set may include false positives), the TAG produces an interest measurement ($\mathbf{f}$) as $\{\mathcal{X}, \mathbf{E}\} \rightarrow \mathbf{f}$ for all images in the database, which allows novel images similar to the example set to be identified. Given the noise inherent to EEG, the TAG's ability to work with noisy labels is critical, as they greatly degrade the performance of many semi-supervised methods [12]. In part, the TAG does this by promoting visual consistency among the example set (prior to generating $\mathbf{f}$) by replacing some example images with more likely candidates from the larger image database (self-tuning) [11], [12].

The suitability of the feature space underlying the TAG graph for differentiating image categories, and the quantity and specific identity of the true and false positive images included in the example image set will affect the TAG's performance. We ran simulations testing the TAG's ability to identify all 62 categories of the testing database as targets of interest when given example image sets of varying quantities of true positive images (i.e. varying precision). For each image category (and each precision level), 50-100 example image sets were randomly selected. TAG interest scores were then computed for all the database images and used to find novel images related to the target category. The quantity of example images was held constant at 20, as several tests with 10, 40, and 60 examples suggested TAG performance was more sensitive to the example set's precision than to its size.

III. RESULTS

A. EEG Classifier Performance

The EEG interest detectors were able to identify target images over distracters. The mean Az score (i.e. the area under the ROC curve) between subjects for the testing RSVP sequences was 0.85 (STD=0.12, N=50), significantly above chance (chance Az=0.5, 1-sided t-test, p<<.001). Similarly, Fig. 2 shows several typical Precision Recall (P-R) curves (one from each subject), which also show how the detectors generally ranked distracters lower than the targets, giving

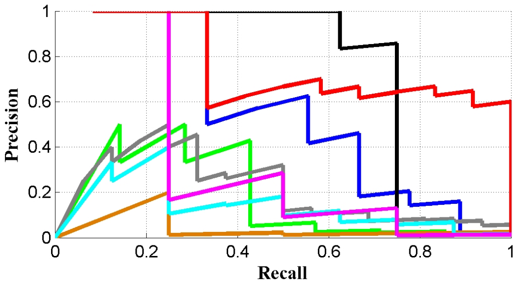

Fig. 2. Typical Precision-Recall curves from each subject.

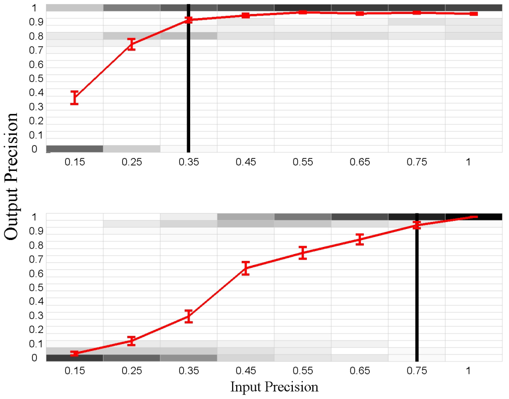

Fig. 3. Examples of TAG performance when identifying target images. Performance varied with both target category and the precision of the TAG's input example set. Dark vertical lines show threshold precisions for ≥90% successful results, red lines show the mean +/- STD.

high precision peaks. However, the large number of distracter images relative to targets meant that similar quantities of target and distracter images were often both given high ranks. Consequently, despite high P-R peaks, using a fixed number of the top-ranked images to create example image sets that show what was of most interest to the user can still lead to significant false positive rates.

B. TAG Classifier Performance

The TAG's ability to identify a target category depended on both the precision of the example image set it was given and the specific target category. Fig. 3 shows the TAG simulation results for two target categories. The target precision between the images with the 20 highest TAG scores are plotted as histograms against the precision of the example image set (darker squares indicate larger numbers of simulations had that result). Larger numbers of true positive examples in the TAG's input typically improved its performance. A bimodality is evident though, with the probability of a very high output precision dramatically increasing for an input precision above some threshold. This example set threshold defines the TAG input requirements needed if the TAG is to successfully identify that type of image as the target. If the TAG identification is considered successful when over half of the top 20 TAG-ranked images are targets, the threshold can be defined as the lowest precision for which the TAG was successful 90% of the time. The bold vertical lines in Fig. 3 illustrate how this threshold varied between image categories. For 18% of the image categories, the TAG had precision thresholds below 0.5 (37% of the categories had thresholds below 1.0). The remaining categories (63%) are likely heavily intertwined with one or more others in the TAG graph, preventing TAG from ever completely disassociating them from the others.

C. Performance Overlap Between Human and Computer Components

How successfully the EEG and CV systems can be directly combined into a single BCI hinges on whether the EEG module can provide example image sets that satisfy the TAG's input precision requirements. Fig. 4 shows the distribution of the TAG's input threshold precisions (defined above) across categories (black, plotted as fraction of categories, precisions >1.0 are assigned to categories for which the TAG never satisfied the success requirement). Overlaid with this is the distribution of target image precision among the 20 highest ranked images from the EEG interest detectors (red, plotted as fraction of RSVP tests, mean=.20, STD=.13, n=50). The region of overlap reflects the subset of categories for which simply using the images with the 20 highest EEG interest scores provides image example sets of adequate precision for the TAG to robustly identify the target category and effectively retrieve relevant novel images from the database.

IV. Discussion

Though the precision of the EEG scores was sufficient to yield robust image retrieval by our computer vision module for many image categories in the test database, the limits in the distributions' overlap in Fig. 4 shows how there were many cases in which the EEG-output/TAG-input requirements were not met. Precisions of the EEG scores are largely constrained by subject variability (both in terms of their EEG quality and their ability to detect targets in the RSVP task) as well as the intrinsically noisy nature of EEG as an electrophysiological measure. The input precision required by the TAG is largely a function of the image database, the features used, and the graphical structure of the

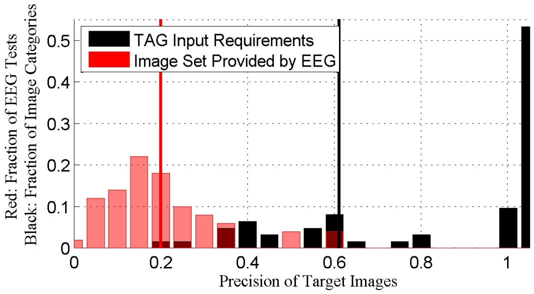

Fig. 4. The overlap between the precision of the EEG interest detector output (red, fraction of 50 RSVPs), and the input precisions needed by the TAG (black, fraction of 62 target categories) determines how effectively the two modules can directly interface. Vertical lines show the means of the EEG (red, 0.20) and TAG (black, 0.61) distributions (for the TAG only categories it captured, i.e. had precisions ≤1.0, were averaged).

model. Improving any of these aspects, whichever is found to be the most practical, would effectively increase the overlap in the input/output requirements of the modules shown in Fig. 4, and improve the full BCI system.

While both modules could be improved, their dual use means that neither must be perfect. Fig. 4 shows that the TAG does not identify many categories regardless of the input example set's precision. This limitation, whether it results from the TAG forgivably confusing categories that are semantically different but really quite similar (e.g. cougar face vs. cougar body), or more significant issues, such as the graph not utilizing a sufficiently rich feature space, must be addressed. However, the TAG does not need to identify these 'difficult' categories when given input sets of very low precision; it simply needs to be able to do so when given sets of a precision obtainable by the EEG component. Similarly, the EEG component does not need to provide example sets of perfect precision; it simply must satisfy the TAG input requirements. For example, Fig. 4 shows that EEG performance does not need to be shifted much to match categories that are reasonably well captured by the TAG. Thus, using entirely new EEG detection algorithms or showing images multiple times for averaging may be unnecessary; and simpler techniques, such as more training for poor performers or using a better method than just taking the top 20 when selecting the images outputted by the EEG component, may suffice.

Another benefit of using both CV and EEG modules is that their mutual interactions can be used to improve overall system performance beyond simply improving the modules independently. For example, it is easier for the TAG to locate all target related images if its example set includes a diverse selection of target images. Similarly, it is often easier for the EEG module to output a higher precision example set with larger numbers of target images in the RSVP. Thus, rather than running both modules once in series (as proposed so far), they could be run in a closed loop fashion, in which the user was exposed to several RSVP sequences, with the TAG being used to gradually increase the prevalence of target images contained in each. This eases the requirements on both the EEG and CV components. In such a closed-loop implementation, the TAG algorithms dedicated to tuning the example set would be deactivated. Thus, rather than TAG trying to reorganize the image database so that many images related to a single category were at the top, a rougher reorganization (requiring a lower input precision) would result. Despite false positive images lingering in the TAG example set, using this reorganization to select the next RSVP sequence would still likely boost the number of target images in the RSVP. Fig. 5a shows that the TAG requires a quite low precision in the example image set (85% of the categories need ≤0.4) when used to gradually increase the prevalence of targets in this way. Fig. 5b shows an example of how the precision of the image sets outputted by the EEG detector improved in such a closed loop implementation (1 subject, 3 target types). In such a BCI, once the number of target images in the RSVP was determined to be of sufficient quantity, the full (i.e. with self-tuning) TAG algorithm would be run to do a final reorganization of the image database.

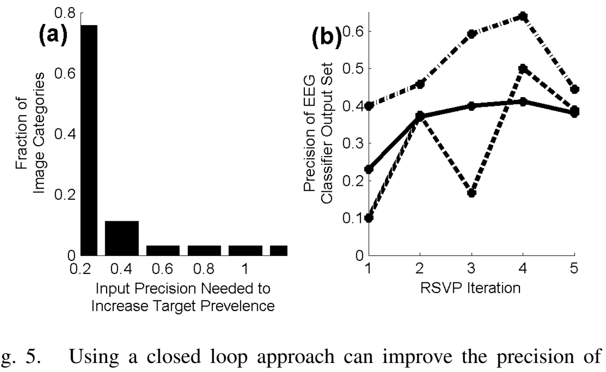

Fig. 5. Using a closed loop approach can improve the precision of the examples provided by the EEG interest detector. (a) Distribution of the lowest example set precisions (across categories) to return more target images in the top 100 TAG results than were in the example set. (b) Increase in precision of images outputted by the EEG detector during 3 closed-loop implementations in which the TAG was used to gradually increase target prevalence in each RSVP sequence.

References

[1] G. Pfurtscheller, G. R. Mller, J. Pfurtscheller, H. J. Gerner, and R. Rupp, "'thought'–control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia." *Neurosci Lett*, vol. 351, no. 1, pp. 33–36, Nov 2003.

[2] J. J. Daly and J. R. Wolpaw, "Brain-computer interfaces in neurological rehabilitation." *Lancet Neurol*, vol. 7, no. 11, pp. 1032–1043, Nov 2008. [Online]. Available: http://dx.doi.org/10.1016/S1474-4422(08)70223-0

[3] J. R. Wolpaw and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans." *Proc Natl Acad Sci U S A*, vol. 101, no. 51, pp. 17 849–17 854, Dec 2004. [Online]. Available: http://dx.doi.org/10.1073/pnas.0403504101

[4] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials." *Electroencephalogr Clin Neurophysiol*, vol. 70, no. 6, pp. 510–523, Dec 1988.

[5] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop CVPRW '04*, 2004, p. 178.

[6] M. C. Potter and E. I. Levy, "Recognition memory for a rapid sequence of pictures." *J Exp Psychol*, vol. 81, no. 1, pp. 10–15, Jul 1969.

[7] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortically-coupled computer vision for rapid image search," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, pp. 174–179, June 2006.

[8] P. Sajda, E. Pohlmeyer, J. Wang, L. C. Parra, C. Christoforou, J. Dmochowski, B. Hanna, C. Bahlmann, M. K. Singh, and S.-F. Chang, "In a blink of an eye and a switch of a transistor: Cortically coupled computer vision," vol. 98, no. 3, pp. 462–478, 2010.

[9] L. C. Parra, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. G. Philiastides, and P. Sajda, "Spatiotemporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks," *IEEE, Signal Processing Magazine*, January 2008.

[10] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694

[11] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang, "Brain state decoding for rapid image retrieval," in *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*. New York, NY, USA: ACM, 2009, pp. 945–954.

[12] J. Wang, Y.-G. Jiang, and S.-F. Chang, "Label diagnosis through self tuning forweb image search," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2009*, 2009, pp. 1390–1397.