Overview of NTCIR-13 NAILS Task

Graham Healy
Dublin City University, Dublin,
Ireland
graham.healy@dcu.ie

Tomas Ward
Maynooth University, Co.
Kildare, Ireland
Tomas.Ward@nuim.ie

Cathal Gurrin
Dublin City University, Dublin,
Ireland
cathal.gurrin@dcu.ie

Alan F. Smeaton
Dublin City University, Dublin,
Ireland
alan.smeaton@dcu.ie

ABSTRACT

In this paper we review the NTCIR-13 NAILS (Neurally Augmented Image Labelling Strategies) pilot task at NTCIR-13. We describe a first-of-its-kind RSVP (Rapid Serial Visual Presentation) - EEG (Electroencephalography) dataset released as part of the NTCIR-13 participation conference and the results of the participating organisations who benchmarked machine-learning strategies against each other using the provided unlabelled test data.

Keywords

Brain-computer Interface, Information Retrieval, Signal Processing, Machine Learning

1. INTRODUCTION

EEG (Electroencephalography) has recently become an accessible method for researchers and users to build and operate BCI (Brain-Computer Interface) applications. While the initial use of such techniques began in clinical / rehabilitative settings for the purposes of augmenting communication and control, a recent trend has been to use such signals and methods in new domains, such as image annotation, which relies on the identification of target brain events to trigger labeling [1][3][5][7]. This trend is particularly relevant to the multimedia IR (Information Retrieval) and HII (Human-information Interaction) communities because in recent years EEG has demonstrated its potential for several applications including annotation of multimedia content, identification of when a user's attention is drawn to something in the real world, or as a source of wearable sensor data to be indexed for later retrieval or analysis.

NTCIR (NII Testbeds and Community for Information access Research) is a conference (18-month schedule) that brings together researchers to develop evaluation methodologies and performance measures for IA (Information Access) technologies. This results in an active research community in which findings based on comparable experimental results are shared and exchanged in an open manner. One topical focus of this is mining knowledge from a large amount of human generated data. NAILS is an affiliated task to support the collaborative evaluation of best-practice strategies for RSVP-EEG image search applications, where researchers benchmark their machine-learning strategies.

In this work, we describe the experimental protocol used to capture the dataset for this task, discuss the motivation behind its construction and outline the results of participating teams at the NAILS pilot task at NTCIR-13.

2. MOTIVATION

Using EEG signals it is possible to detect attention-related events that are understood to be indicative of user interest – or more specifically the allocation of their attention to one particular stimulus as opposed to some other. One characteristic pattern of activity, commonly known as the P300 signal [8], has been a focus of investigation as it can be used as an index of attentional resource allocation to a stimulus such as an attentionally captivating image (due to its infrequency) presented on a screen. This finding has enabled BCI systems to leverage the ability of a user to be able to guide their attention in such a manner so as to be able to provide relevance judgments/ratings on visual stimuli. For example, a user can actively 'look out' for a particular type of image so that when relevant images appear in a high-speed visual presentation sequence known as RSVP (Rapid Serial Visual Presentation)[12], they will subsequently elicit a P300 response that can be detected using signal processing and machine-learning methods. Ultimately this allows the image to be 'neurally' labeled by the participant.

While systems like these have been explored in a proof-of-concept manner in BCI research using a multitude of image-search tasks, the datasets used usually remain unshared between studies, making it difficult to meaningfully compare the machine-learning and feature-processing strategies used, to find those that offer best generalisability both across tasks and participants. EEG responses are rife with variability for numerous reasons, such as differences between experimental participants, between task parameters, or changes that can even occur over the course of an experiment. Such sources of variability impede systematic identification of best-practice methods and strategies in signal processing and machine learning for using neural responses from image presentations to label them. This is what the NAILS dataset seeks to redress, that is to provide a well-constructed test dataset collection to allow researchers to comparatively investigate best-practice strategies for RSVP-EEG image search applications utilizing a range of image-search tasks (in a repeated-measures design).

3. NAILS DATA SET & COLLECTION

3.1 Experimental Task Description

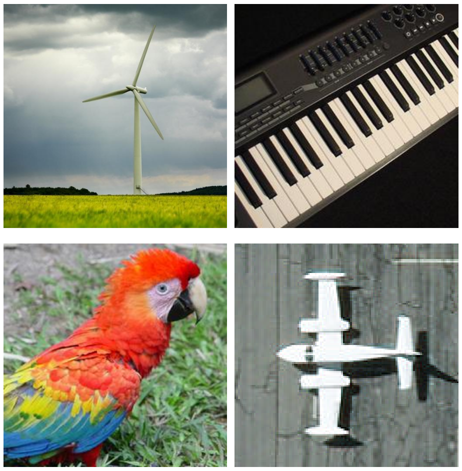

Figure 1: Examples of four target images used in the experiment. In the top left, an example of a wind farm target image like that used in tasks WIND1 and WIND2. In the top right, an example of a keyboard target image used in task INSTR. In the bottom left, an example of a bird (macaw) target used in task BIRD. In the bottom right, an example of an airplane target like that used in UAV1 and UAV2.

Table 1: NAILS Tasks. *standard images were extracted in a balanced manner from the remaining visual categories in the dataset. For the Places365 dataset there were 364 categories remaining and for the VEDAI dataset there were 8 remaining categories.

TaskID	*Dataset*	*Target*	*Standards*
1 - WIND1	Places365	Wind Farm	Field Road
2 - WIND2	Places365	Wind Farm	*All Categories
3 - INSTR	ImageNet	Keyboard	Instruments
4 - BIRD	ImageNet	Macaw	Birds
5 - UAV1	VEDAI	Plane	Pickup
6 - UAV2	VEDAI	Plane	*

Table 2: Counts of training samples per experimental participant (for targets and standards) following trial rejection (not including test set data).

Dataset	*Targets*	*Standards*
101	326	6253
102	140	2316
103	93	2057
104	188	3372
105	356	6748
106	353	6766
107	327	6217
108	193	3691
109	206	4043
110	332	6454

The NAILS dataset collection contains EEG responses to 97,200 images, in total, from 10 experimental participants. Data collection was carried out with approval from Dublin City University's Research Ethics Committee (DCUREC / 2016 / 099). Each participant completed 6 different search tasks (for a particular type of target — see Table 1), where each search task was divided into 9 (approximately 35 second) blocks which were completed in a self-paced manner so as to alleviate strain on participants. In each search task, a participant searched for a known type of target (e.g. an airplane), and was instructed to covertly count occurrences of target images in the RSVP sequence so as to maintain their attention on the task. In Figure 1 we show examples of the target search images used. In each RSVP block, images were presented successively at a rate of 6 Hz with target (search-relevant) images randomly interspersed among standard (non-search relevant) images with a percentage of 5% across all blocks. In each block, 180 images (9 targets/171 standards) were presented in rapid succession on screen. Per participant, there were 486/9234 target/standard examples available. As contaminant eye-movement related activity on the EEG can often contain useful information, epochs (from -1000ms, 2000ms) containing such activity were excluded as they might encourage developed strategies to utilize these non-neural sources of discriminative information. Epochs were filtered to exclude those with a peak-to-peak amplitude greater than 70 µV on EOG and frontal EEG channels. ICA (Independent Component Analysis) was used alongside a wavelet based analysis to confirm that the remaining epochs did not contain non-neural sources of discriminative information. A breakdown of the remaining training data after this process is presented in Table 2. For the NAILS task, this dataset was split into a training/testing set, where 15/285 target/standard trials from each search task (for each participant) were selected to act as a withheld test set in the evaluation.

3.2 Collaborative Evaluation Task Description

Competing teams in the collaborative evaluation using the supplied training data (remaining epochs from blocks not used to extract test set data) were asked to build machine-learning models that maximised the BA (balanced accuracy) score on the withheld testing set (withheld by the NAILS organizers). That means for an evaluation run, a team needed to submit binary predictions for the 18,000 examples given in the test set (900/17100 targets/standards respectively). There were more than 2500/47000 target/standard training examples available across all participants for model training (see Table 2).

3.3 Provided Features / Pre-processing

Three types of preprocessed data were made available to participating organisations: time-series features (time), wavelet magnitude features (w-mean) and wavelet magnitude ratio (w-ratio) features. It was at the discretion of each participating team which combination of these features to use.

Time-series features were extracted from -.5 s to 1.5 s relative to stimulus onset. Each training/testing feature vector contained 3400 numerical values corresponding to 34 channels (incl. HEOG/VEOH) X 100 time-points. Prior to epoch extraction, a band-pass filter (.5Hz to 25Hz) was applied to the raw signal. Time-frequency features were extracted via Morlet wavelets using mne-python [4]. 2 Hz steps between 1 Hz and 11 Hz were used, meaning each wavelet corresponded to one of 6 frequencies (1 Hz, 3 Hz, 5 Hz, 7

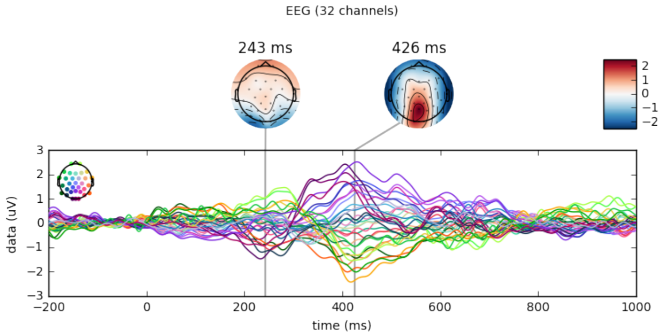

Figure 2: Butterfly plot (ERP averages) of target epochs across all blocks minus average standard epochs across all blocks. Plots are generated using CAR (common average reference). Characteristic P3b activity can be seen at posterior scalp sites approximately between 300ms and 600ms following target detection (peaking at 426ms). The colors on time-series plots indicate electrode location on scalp (upper left).

Hz, 9 Hz and 11 Hz). The number of cycles on each wavelet was set to its respective frequency. To calculate the w-mean feature set, each time-channel-frequency representation was divided into 20 equally spaced time segments over the 2 second period (-.5 s to 1.5 s) for a particular frequency and then the mean magnitude of each of these was calculated. The average magnitude of the 500 ms period pre-stimulus was used to baseline (via subtraction) each epoch. W-ratio features were calculated using the same method as the w-mean features except magnitudes are expressed as a ratio of the average of the baseline period.

3.4 Dataset Validation

In order to validate that the captured data contained useful information for classification prior to sharing the dataset, we applied a basic machine learning analysis using a RBF (Radial Basis Function) kernel SVM (Support Vector Machine) [6]. Each model was trained on a participant-by-participant basis where hyper-parameters (C and gamma) were learned using a randomized grid-search approach. Each model was then applied to the unseen test set data where accuracy measures were calculated (presented in Table 3 and Table 4). A range (and combination) of feature sources (used in baseline approaches) are presented so as to support a better interpretation of the results of participating teams. These baseline results are broken down in both tables under columns 'Time' (which uses the time-series features), 'W-m' (which uses the wavelet mean features), 'W-r' (which uses the wavelet ratio features) and 'T+W' (which uses the time series features and both wavelet feature types). Features derived during the baseline corresponded to the first 1 second of data directly following image presentation (across all channels). In Figure 2, we show a characteristic P3b response acquired from one experimental participant.

These measures, in part verify that the chosen tasks are eliciting the expected characteristic oddball P300 response i.e. it was possible for a participant to do the search tasks as intended. Images tasks were constructed using freely available datasets [13, 10, 9]. These were selected as a good choice given taht they are commonly used datasets with well-researched characteristics that are representative of the visual content typically encountered in multimedia-IR tasks whilst remaining similar to content used in previous RSVP-BCI studies.

Table 3: Balanced accuracy scores for on each participating team's best performing method (ARL17 + QUT) broken down by experimental participant. Baseline results are presented under columns 'Time', 'W-m', 'W-r' and 'T + W'.

Dataset	*ARL17*	*QUT*	*Time*	*W-m*	*W-r*	*T+W*
101	.8219	.7670	.7503	.6006	.5523	.7357
102	.8781	.8512	.8211	.7494	.6465	.8269
103	.8646	.8275	.7664	.6354	.5553	.7213
104	.8877	.8743	.8322	.6845	.6216	.8222
105	.9304	.8921	.8257	.7304	.6725	.8029
106	.8781	.8705	.8026	.6675	.6108	.7956
107	.9170	.8719	.8295	.7047	.6860	.8249
108	.8804	.8658	.8041	.6865	.5933	.8383
109	.8763	.8523	.7930	.6140	.5687	.7518
110	.9041	.8556	.8246	.6482	.6012	.7918
Average	.8839	.8528	.8049	.6721	.6108	.7911

4. PARTICIPATING TEAM'S RESULTS

Two teams submitted an overview paper along with valid predictions to the NAILS Task although nine teams signed up to participate. Team ARL17 was comprised of researchers from *DCS Corporation, Alexandria, VA USA* and from the *U.S. Army Research Laboratory, Aberdeen Proving Ground, MD USA*. The QUT team was comprised of researchers from the *Queensland University of Technology, Brisbane, QLD, Australia* and *Bielefeld University, Bielefeld, Germany*. We describe the results for the approach with the best accuracy for each participating team and compare these to 'naive' baseline approaches.

ARL17 [11] and QUT [2] both made successful submissions whose respective balanced accuracies on the test set are shown in Table 3 and Table 4. Both team's best results respectively achieved balanced accuracy scores on the test set greater than any of the naive baseline approaches. This indicates that both participating team's approaches used a suitably developed strategy i.e. they outperformed a classical off-the-shelf machine-learning strategy like a SVM.

The team ARL17 achieved the highest balanced accuracy of .8839 on their final submission. The winning approach from team ARL17 used within-subject training using a convolution neural network. Other model training strategies such as cross-subject training were explored by the team and found to achieve a lower balanced accuracy (e.g. their first submission achieving a balanced accuracy of .7723). ARL17 submitted five sets of predictions achieving balanced accuracies of .7723, .8459, .8526, .8724 and .8839.

The team QUT submitted two set of predictions. The best performing strategy from the team used a bagging ensemble approach and achieved a balanced accuracy of .8528. Their other approach used a stacked ensemble and achieved a balanced accuracy of .8281.

In Table 3 and Table 4 we present a breakdown across participants and tasks (respectively) of the balanced accu-

Table 4: Balanced accuracy scores for each participating team's best performing method broken down by experimental participant. Baseline results are presented under columns 'Time', 'W-m', 'W-r' and 'T + W'.

Task ID	*ARL*	*QUT*	*Time*	*W-m*	*W-r*	*T+W*
WIND1	.8905	.8609	.8237	.7119	.6435	.8112
WIND2	.8846	.8356	.7746	.6656	.6047	.7821
INSTR	.8114	.7895	.7616	.6367	.5953	.7140
BIRD	.8616	.8191	.7805	.6282	.5837	.7844
UAV1	.9216	.9053	.8381	.6979	.6247	.8200
UAV2	.9335	.9065	.8512	.6925	.6130	.8351
Average	.8839	.8528	.8049	.6721	.6108	.7911

racies achieved by each team's best performing method. In addition, we present four naive classification strategies as a comparison. Three of these used only a single pre-processed data source type while the remaining classification method used features from all three data sources types available (i.e. time, w-mean,w-ratio).

5. CONCLUSIONS

Two teams submitted an overview paper and valid predictions for the NAILS task although nine teams signed up to participate. In the collaborative evaluation, we note that the approaches of both of the teams outperform the baseline strategies that used a SVM classifier. The approach of team ARL17 achieved the highest balanced accuracy overall of .8839. Moreover, their approach also achieved the highest balanced accuracy both when considering the results on a per participant basis and per image search task basis.

Although this was a collaborative evaluation where participating team's machine-learning strategies were ranked in terms of balanced accuracy, it was expected that some signal processing/machine-learning solutions that may perform suboptimally to others in terms of accuracy alone may offer other advantages in terms of speed, model complexity, neurophysiological interpretability and/or cross-task/user applicability. Both active teams achieved these aims, with ARL17 testing a variety of machine-learning approaches using different training strategies and QUT via its exploration and visual summarization of channel selection as part of the model training process.

Very often, useful features for ERP-driven BCI interfaces leverage time-domain information in the signals due to the time-locked nature of the P300 response. One team (QUT) directly used the provided wavelet features as part of their approach while the ARL17 team used time-frequency related features via the involvement of temporal convolutions as part of their solution.

In this paper we have described the creation of the NAILS dataset, including the motivation behind the challenge and an account of key details in its construction. We also described important parameters of the dataset such as those available from the prior validation tasks carried out. Finally as a demonstration of the evaluation value of the NAILS dataset, we summarize the results of participating teams in the associated NTCIR-13 NAILS challenge.

Acknowledgments

This research was supported by Science Foundation Ireland under grant number SFI/12/RC/2289.

References

[1] G. Healy and A. F. Smeaton. Eye fixation related potentials in a target search task. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4203–4206, Aug 2011.

[2] H. Hutson, S. Geva, and P. Cimiano. Ensemble Methods for the NTCIR-13 NAILS Task. *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, Tokyo, Japan, 5-8 December 2017.*, 2017.

[3] S. Koelstra, C. MÃijhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis using physiological signals. 3(1):18–31, 2012. eemcs-eprint-21368.

[4] E. Larson, A. Gramfort, D. A. Engemann, jaeilepp, C. Brodbeck, M. Jas, T. L. Brooks, jona sassenhagen, and M. L. et al. mne-tools/mne-python: v0.15, Nov. 2017.

[5] A. R. Marathe, A. J. Ries, V. J. Lawhern, B. J. Lance, J. Touryan, K. McDowell, and H. Cecotti. The effect of target and non-target similarity on neural classification performance: a boost from confidence. *Frontiers in Neuroscience*, 9:270, 2015.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S.-F. Chang, and P. Sajda. Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases. *Journal of neural engineering*, 8 3:036025, 2011.

[8] J. Polich. Updating P300: an integrative theory of P3a and P3b. *Clin Neurophysiol*, 118(10):2128–2148, Oct 2007.

[9] S. Razakarivony and F. Jurie. Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187 – 203, 2016.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[11] A. J. Solon, S. M. Gordon, B. J. Lance, and V. J. Lawhern. Deep Learning Approaches for P300 Classification in Image Triage: Applications to the NAILS Task. *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-13, Tokyo, Japan, 5-8 December 2017.*, 2017.

[12] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520 EP –, Jun 1996.

[13] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016.