# Exploring EEG for Object Detection and Retrieval

Eva Mohedano[1], Amaia Salvador[2], Sergi Porta[2], Xavier Giró-i-Nieto[2],
Kevin McGuinness[1], Graham Healy[1], Noel E. O'Connor[1] and Alan F. Smeaton[1]

[1]Insight Center for Data Analytics, Dublin City University, Dublin, Ireland
[2]Universitat Politècnica de Catalunya, Barcelona, Catalonia/Spain

## ABSTRACT

This paper explores the potential for using Brain Computer Interfaces (BCI) as a relevance feedback mechanism in content-based image retrieval. Several experiments are performed using a rapid serial visual presentation (RSVP) of images at different rates (5Hz and 10Hz) on 8 users with different degrees of familiarization with BCI and the dataset. We compare the feedback from the BCI and mouse-based interfaces in a subset of TRECVid images, finding that, when users have limited time to annotate the images, both interfaces are comparable in performance. Comparing our best users in a retrieval task, we found that EEG-based relevance feedback can outperform mouse-based feedback.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; H.5.2 [**User Interfaces**]: Input devices and strategies

## General Terms

Experimentation, Design

## Keywords

Brain-computer interfaces, Electroencephalography, Rapid Serial Visual Presentation, Classification, Instance Retrieval

## 1. MOTIVATION

The exponential growth of visual content and its huge diversity has motivated considerable research on how documents can be retrieved according to user intentions when formulating a query.

Advances in image processing and computer vision have provided tools for a perceptual and semantic interpretation of both the query and the indexed content. This has allowed the development of retrieval systems capable of processing queries by example and concepts.

The role of a human user during visual retrieval is critical, and his judgment about the correctness of the retrieved results can greatly speed up the search processes. This kind of relevance feedback has been demonstrated to significantly improve retrieval performance in image [10] and video [1] retrieval. Manually annotating images using a mouse, especially in a visual retrieval context can be tedious and mentally exhausting. In that such a scenario, EEG-based brain computer interfaces offer a potential solution as a mechanism to quickly annotate images.

## 2. RELATED WORK

EEG signals have been used for object detection in [2], where authors aim to detect airplanes in a dataset of satellite images from the city of London. The work in [3] expands the catalog of objects in very simple images, where the object on a black background occupies the whole image.

EEG signals have been also used for image retrieval in [9], where authors used EEG relevance annotations to retrieve specific concepts in a complex dataset of keyframes from TRECVid 2005. However, while that work aimed at detecting concepts depicted by the whole image, we focus on the more challenging task of detecting a local object in a complex scenario. Another similar work [8], addresses the usage of EEG for image retrieval by formulating the problem as a semi-supervised learning problem. The noisy relevance labels generated by EEG classifiers on a small subset of a large database are expanded through a visual similarity graph. In our experiments, the highest and lowest EEG relevance scores are used to train a binary classifier, instead of feeding a semi-supervised algorithm. This way results are more comparable to the binary relevance labels collected with the mouse interface.

## 3. EXPERIMENTAL SET-UP

Following previous work on EEG-based image retrieval [3, 4, 8], the experimental design was based in the 'oddball paradigm', in which two different stimuli are presented to the user in random order and with different probabilities. One stimulus (the 'target') appears with low probability during the visual presentation and the other (the 'distractor') appears frequently. Users are asked to focus on detecting the target and to express their reaction by either counting or pressing a button.

In this context, it has been found that when a user reacts to a target stimulus, a P300 wave appears in the captured brain signals. This wave is a kind of Event Related Potential (ERP) and it consists in a positive peak in the EEG

*ICMR'15,* June 23–26, 2015, Shanghai, China.
Copyright © 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00.
DOI: http://dx.doi.org/10.1145/2671188.2749368.

activity around the 250ms-500ms range after the stimulus presentation.

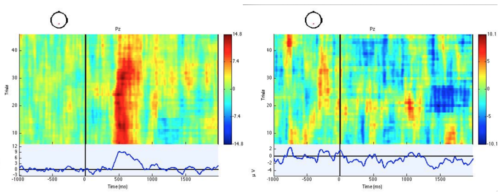

Figure 1: Visual evidence of discriminative P300 response on averaged epochs for 1000 images (5 blocks of 200) captured at 5Hz for the Pz channel. On the left, the response on target images. On the right, the response for 50 distractors.

For our experiments, we selected a 'real world' set of images from a subset of the TRECVid 2013 instance search dataset. Three queries were selected, each one containing 4 visual examples. Figure 2 shows one visual example of each of the three selected queries. Using the ground truth labels provided by TRECVid, 1000 images were selected for each query. To adapt the data to the oddball paradigm, the 'target' ratio of the dataset has been set to 5% (i.e. for each query, 50 relevant images and 950 non-relevant images are used).

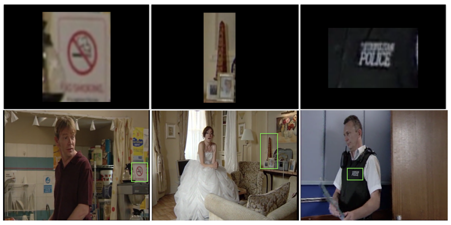

Figure 2: Visual examples of the three query objects (top) and the actual images where they appear (bottom). From left to right: 1. A circular 'no smoking' logo, 2. A small red obelisk, 3. A Metropolitan Police logo.

The selected images are presented in a Rapid Serial Visual Presentation (RSVP). For each query, the 1000 images are divided in 5 blocks of 200 images keeping a 5% target ratio within each block (i.e. 10 'target' images and 190 'distractors'). Once this division is done, the images are randomly sorted within the block. These blocks are presented to the user one after the other, allowing him/her to rest for a few seconds between blocks. The user's task consisted in counting/pressing a button every time that they detected a target image during the RSVP. The visualization rates that have been tested in our experiments are 5Hz and 10Hz, therefore the duration of the experiment for a single query is 200 and 100 seconds.

Eight volunteers between the ages of 19 and 33 participated in the experiments: six women and two men. Of the eight users that participated in the task, two completed the experiment with a RSVP at 10Hz and four at 5Hz. The last two users completed the task twice, once at 10Hz and once at 5Hz.

The BCI device used was a 32 channel actiCHamp amplifier, which was connected to the user locating the electrodes according to the 10-20 system.

## 4. EEG SIGNAL PROCESSING

The signals captured by each one of the 32 sensors were re-referenced to the average of all the channels. Then, the sample rate was reduced from the original 1000Hz to 250Hz and the signals were band-pass filtered from 0.1Hz to 20Hz. The epochs related to each visual stimulus were extracted, obtaining 1000 epochs with EEG activity from 1 second before to 2 seconds after each image. Then, for each of the 32 signals, the period from 200ms to 1s was taken as the discriminant time region to discern between EEG responses of targets and distractors (see Figure 1) and with the sample rate reduced to 20Hz, generating a 16 sample vector per channel. Each of the 16 samples per channel was the result of computing the average of 24 samples windows with 50% of overlap between each other, which we found was a better strategy than to just decimate the signal. Finally, we build a single feature vector for the image as the concatenation of the 32 channels, generating a 512-dimension vector per image.

We used a linear SVM model with default parameters to classify the EEG signals. Every user had his/her own models, i.e. for each user, the data associated to the presentation of 2 queries (2000 EEG epochs) was used as training examples, and the remaining query was used for testing. This procedure was repeated 3 times to obtain a classification prediction for the images of the three queries.

We used the Area Under the Curve (AUC) of the Receiver Operating Characteristic space (ROC) to evaluate the performance of the models.

## 5. OBJECT DETECTION

After generating a relevance score by using the SVM models, the images were ranked from higher to lower relevance score. The top row of figure 3 shows a few examples of some of the correctly classified images for one of the target objects. In these examples we can see images that have been correctly detected as relevant despite of the fact that, in most of them, the object appears to be very small (sometimes even incomplete). Such behavior suggests that users do not only recognize the object itself, but also its context.

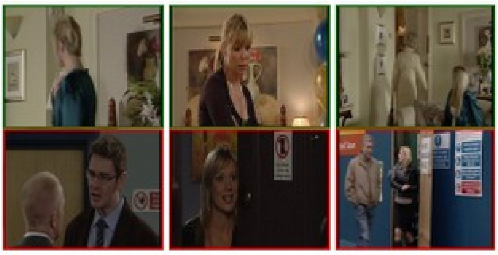

Figure 3: Images with high SVM classifier scores for user A and and query 2 (top) and images with high SVM classifier scores (yet not relevant) for user D and query 1 (bottom).

The bottom row of figure 3 shows some sample images that obtained high SVM scores for query 1, despite not being relevant to the query. It is interesting to see that all these images contain objects that are similar to the queried object. This also suggests that users respond not only to the general appearance of the image but also to the object they are looking for.

## 5.1 User diversity

Figure 4 shows the ROC curves for all the users who participated in the experiment. Due to the the huge diversity in user's performance, we define two user profiles:

- The *expert* user, who is familiar with the presented images (they have seen the images in advance) and with the purposes of the experiment.
- The *novice* user, who has had no previous exposure to the images.

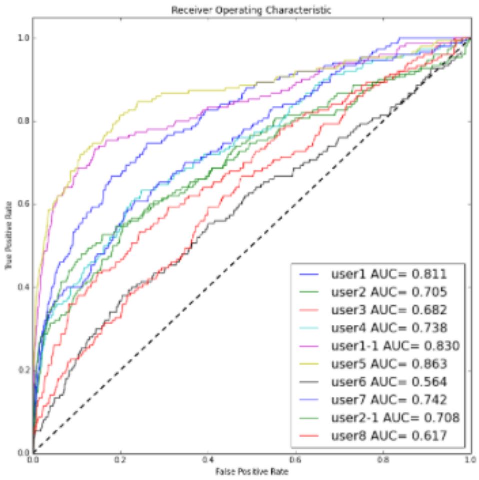

**Figure 4: ROC curves for the 8 users. Users have as many curves as the number of experiments they performed.**

Following the criteria stated above, 3 of our users were selected as experts and 5 of them as novice users. We then average the AUC values of the three queries for expert and novice users separately, reaching values of 0.758 and 0.642, respectively. The average AUC achieved by expert users is significantly higher than the one for novice users (t-test, p=0.005317, sample size = 9). This is a reasonable result, since expert users are already familiar with the images look like prior to the experiment, which gives them an advantage.

## 5.2 Visualization rates: 5Hz vs 10Hz

The second experiment explores the impact of different visualization rates in users' performance. We compare the performance achieved by those users who did the experiment at 10Hz with those who did it at 5Hz. We average the AUC values of the three queries for 5 and 10 Hz experiments separately, obtaining values of 0.775 and 0.734, respectively. This difference between the two is, however, not significant (t-test, p = 0.1397, sample size = 12), which leads us to conclude that 10Hz is a reasonable visualization rate for users to be able to identify objects in complex images and, more importantly, reduces the length of the experiment by half.

# 6. EEG VS. MOUSE FOR RETRIEVAL

In order to compare EEG- with mouse-based annotations, we use the interface from [6] on the same set of images described in Section 3. This interface displays the images in a thumbnail grid in the same order as in the EEG experiment. The time given to the users to annotate the images with the mouse interface has been restricted to the same amount of time that users spend visualizing images in the EEG setup, i.e. 200 seconds for the 5Hz configuration and 100 seconds for the 10 Hz one.

## 6.1 Retrieval within our dataset

The annotations obtained with both EEG and mouse-based mechanisms were used to sort the 1000 images for each query and produce a ranking.

The ranking for the mouse setup is constructed as follows: Given a set of positive annotations and their time stamps, we define two sets of images $p_a$ and $n_a$, where $p_a$ are the positive annotations themselves and $n_a$ contains all those unmarked images presented before the last positive annotation (i.e. we assume that all the observed images that have not been clicked are negative). Then, the ranking is built to ensure that the images in $p_a$ are always on top and the images in $n_a$ are always at the bottom. The remaining set of images are placed in between in the same order in which they were displayed in the mouse interface.

The ranking for EEG is constructed by just sorting all the images by their SVM score in descending order.

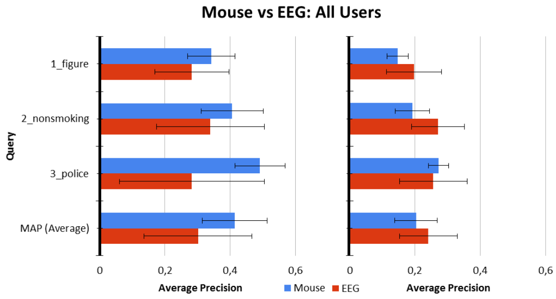

**Figure 5: Mouse vs EEG. On the left, averaged values for all Users at 5Hz / 200 seconds. On the right, for 10Hz / 100 seconds.**

We can see in Figure 5 that there is an accuracy drop in both Mouse and EEG when time is reduced from 200 to 100. Interestingly, this drop in performance is a lot higher for the Mouse-based mechanism. The same time limitation on the EEG approach does not seem to significantly affect the final accuracy, which points out the potential of EEG signals, specially when the images are displayed at high frequency rates.

## 6.2 Retrieval in a larger dataset

This last experiment consists of combining the collected annotations with visual descriptors to retrieve images from a bigger dataset. The chosen dataset consists of 23,614 frames selected from the TRECVid Instance Search dataset from 2013. We extract visual descriptors for all our images from a Convolutional Neural Network (CNN) pre-trained with ImageNet [7] using the *Caffe* software [5]. The selected feature vector is the output of Layer 7 (the second top fully connected layer), which has 4,096 dimensions.

We then train a linear SVM model with default parameters for each query from the obtained annotations. This model is used to score to all images in the larger dataset and then sorted in descending order.

In order to train the model it is necessary to select positive and negative examples. For the mouse-based interface, we consider as positive examples all the images clicked by the users. Negative examples are those observed images that the user saw but did not click during the annotation time. In the EEG case, the annotations are selected according to the confidence scores given by the EEG models. We sort the images in descending order and take the top 10 and bottom 100 images as positive and negative examples, respectively.

The mean average precision values considering three queries and users for the mouse and EEG systems are 0.27 and 0.21 respectively. Results for the mouse configuration are significantly higher, but these results are averaged across all the users, and we know that there is a high variability among them (Section 5.1) and that this variability is especially present in the EEG case.

To make a comparison that is not subject to user diversity, Figure 6 shows the performance of the best user for both interfaces, in the two time configurations. As expected, when the user interaction time is set to 200s (5Hz RSVP), the mouse-based annotations are more effective than the EEG ones, obtaining mean AP of 0.49 vs 0.27, respectively. Nevertheless, when we consider the best users for the the 100s configuration (10Hz RSVP), we obtain a similar performance for the two, with EEG a slightly better on average with a mAP value of 0.37 against a 0.32 for mouse-based feedback.

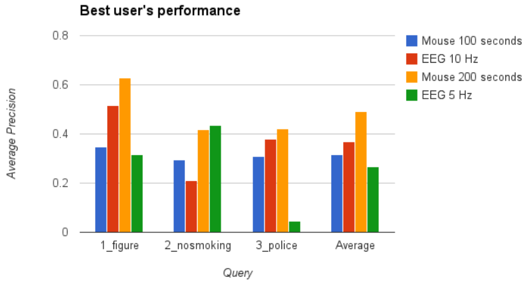

**Figure 6: Average Precision comparison between mouse and EEG annotations for the best user in both configurations**

## 7. CONCLUSION

In this paper, we have presented and studied the potential application of EEG as a mechanism towards relevance feedback and compared it to the traditional "click-based" one for an object retrieval task. We have compared the EEG-based relevance feedback signal with standard "click-based" relevance feedback from a mouse and find that comparable accuracy can be achieved with the EEG-based approach, especially for those users who are familiar with both the brain computer interface and the image dataset. We have seen that there is potential for EEG specially when using higher RSVP rates, which allow for faster annotation with little drop in performance.

## Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289, the project BigGraph TEC2013-43935-R, funded by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF), as well as the donation of a GeoForce GTX Titan Z from NVIDIA Corporation.

## 8. REFERENCES

[1] A. Amir, M. Berg, and H. Permuter. Mutual relevance feedback for multimodal query formulation in video retrieval. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '05, pages 17–24, 2005.

[2] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig. Brain activity-based image classification from rapid serial visual presentation. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 16(5):432–441, 2008.

[3] G. Healy and A. F. Smeaton. Optimising the number of channels in eeg-augmented image search. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, pages 157–162. British Computer Society, 2011.

[4] Y. Huang, D. Erdogmus, M. Pavel, S. Mathan, and K. E. Hild II. A framework for rapid visual image search using single-trial brain evoked responses. *Neurocomputing*, 74(12):2041–2051, 2011.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

[6] K. McGuinness, E. Mohedano, Z. Zhang, F. Hu, R. Abatal, C. Gurrin, N. O'Connor, A. F. Smeaton, A. Salvador Aguilera, X. Giró Nieto, et al. Insight centre for data analytics (dcu) at trecvid 2014: instance search and semantic indexing tasks. 2014.

[7] 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, 2014.

[8] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang. Brain state decoding for rapid image retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 945–954. ACM, 2009.

[9] J. Yang. *A general framework for classifier adaptation and its applications in multimedia*. PhD thesis, Carnegie Mellon University, 2009.

[10] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*, 8(6):536–544, 2003.