Informed Perspectives on Human Annotation using Neural Signals

Graham F. Healy[1], Cathal Gurrin[1], and Alan F. Smeaton[1]

Insight Centre for Data Analytics,
Dublin City University, Glasnevin, Ireland
me@grahamhealy.com

Abstract. In this work we explore how neurophysiological correlates related to attention and perception can be used to better understand the image-annotation task. We explore the nature of the highly variable labelling data often seen across annotators. Our results indicate potential issues with regard to 'how well' a person manually annotates images and variability across annotators. We propose such issues arise in part as a result of subjectively interpretable instructions that may fail to elicit similar labelling behaviours and decision thresholds across participants. We find instances where an individual's annotations differ from a group consensus, even though their EEG (Electroencephalography) signals indicate in fact they were likely in consensus with the group. We offer a new perspective on how EEG can be incorporated in an annotation task to reveal information not readily captured using manual annotations alone. As crowd-sourcing resources become more readily available for annotation tasks one can reconsider the quality of such annotations. Furthermore, with the availability of consumer EEG hardware, we speculate that we are approaching a point where it may be feasible to better harness an annotators time and decisions by examining neural responses as part of the process. In this regard, we examine strategies to deal with inter-annotator sources of noise and correlation that can be used to understand the relationship between annotators at a neural level.

Keywords: Brain-computer interface, EEG, hci, information retrieval, semantic

1 Introduction

In recent years there has been a focus in the multimedia analytics community towards extracting semantically meaningful value from multimedia content. Mining the semantic content of multimedia data to extract visual features is an application of content-based image retrieval (CBIR) [7]. In a naive implementation, low-level image features such as colour, texture, shape, local features or their combination could represent images [2]. However, it was noticed that the problem of the Semantic Gap arises where an individual's interpretation of multimedia content can be different from that of a machine. As a consequence in recent years higher-level semantic extraction (typically based on deep

learning) has become popular [6]. In order to be effective, such deep learning approaches require a significant amount of training data, which naturally poses a large human-factor overhead in terms of the selection of appropriate training data.

One convenient solution that has gained favour is the integration of non-expert annotations via a process of crowd-sourcing, in which contributions are solicited from a large group of people, usually from an online community. Services such as Amazons Mechanical Turk [13] provide the facility to support this. Crowd-sourcing has been applied to generate a variety of multimedia analytics datasets (and subsequently tools) such as food labelling [9], machine translation [1] and concept detection in digital images [6]. The widespread adoption of crowd-sourcing has substantially benefited the multimedia analytics community, yet it relies on a human component that is not well controlled or even well understood. Jia et al. [6] ask how can one trust the labels obtained from such services? They propose an algorithm that reduces the number of labels required, and thus the total cost of labelling, while keeping error rates low on a variety of datasets.

What is the clear from such trends is the important need for human annotators although they are potentially unreliable and in many instances agreement between non-expert annotators can be low. Since it can be difficult to solely interpret the factors that might be affecting the way people annotate from the annotation data and/or post-task questionnaires, we propose the use of a neurally combined perspective. For multimedia analytics it is important to have good machine generated annotations but in order to do this we need good annotations and thus good annotators.

Given a specified data annotation task, we explored relationships between behavioural (manual annotations) and neurally derived predictions from an EEG (Electroencephalography) device, for eight annotators annotating a image collection for the presence of semantic concepts. Our findings suggest that many annotators who partake in the activity may do so in a careless manner, neglecting to annotate images of interest. We can make this observation based on the analysis of the EEG signals understood to correlate with attentional-orientation processes (i.e. *interest*). In the subsequent sections we describe our experimental procedure, discuss the results of our experiment, and frame our findings in terms of future perspectives on the application of EEG in understanding annotation tasks as they can be performed on a crowd-sourced level. We end this paper by making a list of suggestions that could be employed in future annotation activities to reduce the risk of incomplete or incorrect annotations being generated.

2 Background to EEG in Annotation Tasks

The P300 ERP (Event-Related Potential) is a neural signal present in response to stimuli (such as images) that significantly capture or engage a participant's attention [10]. While there are a large number of measures that can be extracted

from EEG signals, the 'P300' (P3) is typically understood to provide a measure of attentional allocation/orientation to a stimulus and subsequently has been the focus of a wide variety of BCI (Brain-computer Interface) application research including EEG-image labelling tasks [3], [8], [4], [5]. Although similar approaches of using EEG for image/stimulus annotations tasks have been explored by others before, we extend upon this prior work by examining how these techniques can allow us to better understand annotators and potential underlying factors relating to quality of the annotations.

In this work we explore an alternative and complementary approach by integrating EEG (Electroencephalography) to derive multimedia dataset annotations in a controlled laboratory environment. We do this by examining neural signals present at the time a stimulus was first seen (and subsequently encoded) with respect to later annotations provided by the user/annotator as to which particular visual stimuli significantly captured their attention. We propose that we should find P3-related activity that should ultimately later correspond with the ground-truth labelled data from the annotator captured later in the experiment.

We use Lifelog images as our dataset because our research interest is directed at understanding and decoding neural responses to semantically rich imagesets i.e. in this task, places and artefacts that the participant would be familiar with. Although in this work we shape our focus and approach around attention orientation related processes, other authors have used implicit responses and information contained in other ERP time windows that are specific to other neural processes such as face perception/processing for instance [11].

3 User Annotation Experiment

In order to assess the level to which ERP responses could be used in an image annotation task (and understand the information they provide about how a person completes the annotation task) eight participants were recruited from the research staff/postgraduate body. Each participant took part in three phases of experimentation, during a single session. In the first phase (the 'pay-attention' phase) participants viewed previously unseen lifelog images without a specific task to accomplish; this is to ensure that particular images that capture their attention are not as a result (e.g. semantic/visual similarity) of target image categories they will be seeking during the model-training phase (phase II). In the second phase (model-training phase) the annotator is shown images and they must identify specific objects of interest; in this phase we already have the ground truth from the attention phase. We used this second phase to build prediction models (from EEG/ERP signals) and then applied these on phase I data. Finally in phase III (annotation phase), the annotators annotate the images they seen in phase I so that we can better understand the relationship between their explicit (manual) and their more implicit-like EEG-based annotations.

The eight participants recruited were academic researchers working in a computer science department; hence they would understand the importance of

the annotation process and the need for accurate annotations to be provided. Specifics of the experiment such as the reasoning for the ordering of tasks was not explained to participants in advance, only that they would be performing target search experiments that involved different attention tasks and providing feedback by behavioural (key press) response.

3.1 EEG Setup and Configuration

The EEG data was recorded using an ActiCHamp 32-channel EEG system at 10-20 electrode locations Fp1, Fz, F3, F7, FT9, FC5, FC1, C3, T7, TP9, CP5, CP1, Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, CP6, CP2, Cz, C4, T8, FT10, FC6, FC2, F4, F8 and Fp2. Signals were bandpassed between .2 Hz to 20 Hz and epochs of 50 ms to 750 ms relative to stimulus onset were extracted. Epochs were baselined to the 200 ms directly prior to stimulus presentation. Impedance was kept below 5 kOhm across channels. Noisy channels were removed and interpolated. No trial rejection strategy was used. EEG signals were rereferenced to common average reference. Independent Component Analysis was used to removed activity related to eye movements such as blinks.

Features were extracted from epoch windows generated on averaged clusters of localised channels corresponding to typical topographic mappings of related ERP phenomena: (O1 + O2 + Oz), (P3 + P7 + O1), (P4 + P8 + O2), (Pz + CP1 + CP2 + Cz), (Cz + C3 + C4 + CP1 + CP2 + FC1 + FC2), (Cz + Fz + FC1 + FC2), (Fz + Fp1 + Fp2), (F3 + F4 + Fz + FC1 + FC2), (F3 + F7 + Fp1), (F4 + F8 + Fp2), (C3 + T7 + FC5 + CP5) and (C4 + T8 + FC6 + CP6). This yielded 12 (pseudo-) channels of 50 features per channel (600 features per trial). All of these considerations were based on our experience of using EEG devices over a number of years for BCI-related applications.

3.2 Experiment Outline

Participants were seated approximately 60cm from the computer screen and given an overview of the nature of the task (i.e. high speed image search). Following this introduction, they begin the experimental session. Participants were asked to refrain from physical movement to avoid any chance of missing an image and also to reduce movement-related noise in the EEG recording. The total experiment took about approximately 50 minutes per participant (including setup time). The purpose of the experiment was not explained to participants until they had completed all 3 phases of experiments.

Phase I - Pay-Attention Phase. In the first phase of the experiment participants viewed a stream of lifelog images captured from the perspective of a person walking through Dublin city centre; the participants were seeking to identify images that captured their attention. There were 461 images chosen from an archive of approximately 1800 images after filtering. These images were manually

filtered for those containing occluded images by clothing and such. Examples of the lifelog images employed are shown in Figure 1.

These 461 images (consistent across participants) were presented in a randomised order each time to participants via a RSVP (Rapid Serial Visual Presentation) protocol at 5Hz (high speed). Participants were instructed to look out for images that captured their interest/attention. This phase was completed first so as to minimize any carry-over effects from image content to be used in the training of phase II, such as attention being orientated towards a particular image not because it is of interest itself but that it might appear to be related to an image (visually/semantically) used in the training block. In effect as this block is completed first across all participants, the application of our later trained classification models (from phase II) to this block are not compromised by non-related attentional artefacts as a result of recognising images potentially related to the training blocks.

Phase II - Model Training Phase. In phase II we collect data to be able to train EEG-prediction models that can be used to annotate the images from phase I. In each of these blocks a participant was required to count the collective number of occurrences of 4 predefined target concepts/categories. Prior to the start of phase II each participant was shown target images for these concepts and allowed time to become familiar with them before they felt they could accurately recognise them at a 5 Hz presentation speed.

Each target category was of a recognisable building/object which could be easily recognised from different camera angles/views. Each target category had 2 related images (e.g. 2 different photos containing the same building/artefact). Target categories were selected to be different across participants but balanced across 10 possible categories across participants. That is each participant had a unique combination of targets to search for compared to other participants with some overlap.

During this phase, 4 training blocks (80 seconds each) were completed containing 80/300 target/non-targets (totalling 320/1200 target/non-targets), providing a means of capturing ground truth labelled data that in turn could be used to train EEG-prediction models. Each block contained an equal number of each of the 4 target concepts to be searched for. Participants were required to count occurrences of target concepts so as to tune their attention to the task and validate the participant was capable of performing the task. Target categories used were images of: 1. Starbucks shop, 2. a pub (The Earl), 3. Temple Bar, 4. a restaurant (Bull & Castle), 5. a government building, 6. a game shop, 7. a pub (Grogan's), 8. the outside of a public shopping mall, 9. a cafe (Bewley's) and 10. a statue in a public park.

Following completion of each training block, total counts were reported by the participants. The requirement to count targets in this way during a RSVP visual search task is known to mitigate issues of wandering attention during a task. For our classification model training we used data available over all 4 training blocks.

A Bayesian Ridge Regression classifier was used to train a model for each participant using the training data available in blocks from phase II (4 blocks of 80/280 target/non-targets). A randomised cross-validation grid search was used for parameter tuning of the model with the final accuracy of the chosen parameters validated on an independent subset of the available training data. Each model's 4 hyper-parameters (lambda 1, lambda 2, alpha 1, alpha 2) were sampled from a log space distribution of -20,20 (base 10) with 10 values per hyper-parameter. 20 iterations (random parameter selection) were used in estimating optimal parameters for each participant's model. A withheld test set from the training (and parameter selection) process (stratified) of 20% (64/224 targets/non-targets) was used to ascertain the EEG classifier model's accuracy on training data as shown in Table 1 (Training Accuracy).

The model trained on blocks available from phase II was then applied to the EEG data captured in phase I to derive 'interest' predictions for each image. ROC (Receiver operating characteristic)-AUC(Area Under Curve) metric was used to measure accuracy.

Phase III - Annotation Phase Following completion of the training/calibration blocks participants were asked to label (via keystrokes) the 461 images shown in phase I into 3 categories: (A) I do not remember this image from the first block, (B) I remember this image but there was nothing particularly interesting about it that captured my attention and (C) I remember this image capturing my attention in a significant way. This was done so that we could evaluate the agreement between predictions from the phase II EEG-based model when applied to phase I, to the participant's own manual annotations of the phase I image content (provided in phase III). These labels were later remapped to new labels hereafter referred to as targets (C) and non-targets (A & B) for the rest of text.

In Table 1 we show the counts for targets across annotators and the (model) accuracy of the phase 2 learned EEG model applied to phase 1 data compared to participant's annotations that were acquired in phase III. Accuracy is measured using ROC-AUC so as to allow comparison between varying target/non-target counts across participants. Mann-Whitney U p-values are also presented to assist in assessing accuracy.

We refer to EEG-prediction scores for the remainder of this text as the respective annotator's phase II learned model applied to their phase I EEG data.

4 Results of the User Annotation Experiment

In Table 1 it can be seen that, for each participant, it was possible to discriminate target from non-target concepts with high accuracy for phase II data. Surprisingly, however, the accuracy of the phase II model when applied to phase I data using participant's manual annotations as ground truth labels (captured in phase III) dramatically decreases (measured via ROC-AUC). A bootstrapped ROC-AUC statistic reveals that for 3 of these participants (annotators 1,3 and

(A) Manual annotations

(B) EEG-based model

Fig. 1: Examples of top-ranked images across participants using manual annotations (A) and EEG-based annotations (B)

5) the model does perform above a chance level (α <.05). As ROC-AUC might not be a sensitive/suitable enough as a metric we use a Mann-Whitney U test that also reveals a similar pattern of relationships (also shown in Table 1). There is indication here too that annotators 2 & 6 might also display an albeit weak relationship between their EEG-predictions and behavioural responses.

In order to help understand the underlying relationships present between annotations and EEG predictions across annotators, in Figure 3 we show univariate Mann-Whitney U tests (a measure we use for cross prediction accuracy) for each annotator's labels predicting each other person's EEG-based scores. Here, we can see for instance EEG-predictions for annotators 4, 7 and 8 are not well predicted by other annotator's labels (and nor the annotator's own labels). Conversely, we can see for instance labelling data from annotators 7 & 8 seemingly predict other annotator's EEG-predictions scores better than their own EEG-prediction scores. Such evidence indicates the presence of shared target image relationships across annotators. Comparatively, in Figure 4 we can see underlying patterns of correlated (via Spearman's rho) EEG-prediction scores across annotators for phase I images. The presence of such relationships is further confirming shared responses across annotators irrespective of how they later label. A PCA (Principal Component Analysis) of annotator EEG-prediction scores across images in phase I indicate via bootstrapping statistic (testing % variance accounted for in first component against distribution generated using randomised image - prediction score mappings within annotators) reveals PCA is finding significant patterns of co-varying activity between annotator's EEG-predictions (p <.05).

Examining Spearman's rho correlation on the consensus annotation counts (from phase III) and average EEG-prediction scores (from phase II trained model applied to phase I data) across all annotators we find a near significant correlation (Spearman's rho = 0.087, α = 0.06, N=461). When only data from annota-

Table 1: * indicates significant results @ α <.05 (via bootstrap resampling randomization test). # Targets: Number of targets (i.e. 'interesting' images) labelled per annotator during phase III. Model Accuracy: ROC-AUC accuracy of phase II model applied to phase I EEG-data with respect to labels obtained in phase III. Training Accuracy: ROC-AUC of phase II models applied to an independent test set of data kept in phase II i.e. not part of parameter selection or training. MW-U: P-Value (Mann-Whitney U) comparing EEG-prediction scores (obtained from phase II prediction models applied to phase I data) for targets v non-targets for each annotator (as obtained in phase III).

ID	# Targets ('Interesting')	Model Accuracy	Training Accuracy	Bootstrap AUC	MW-U
1	71	.64 *	.92	.56	<.00
2	19	.41	.92	.61	.042
3	12	.64 *	.85	.61	.043
4	73	.53	.90	.57	.178
5	27	.62 *	.91	.59	.009
6	20	.54	.93	.62	.035
7	122	.45	.85	.55	.267
8	34	.49	.89	.57	.157

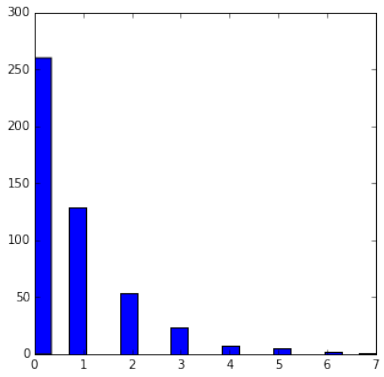

Fig. 2: Histogram showing distribution of summed manual annotation scores across all annotators for all images (N=461). For example a score of 6 indicates 6 annotators marked the image as a target.

tors 1,3 and 5 are examined in this way we find greater correlation (spearman's rho = 0.204 α <.000001 N=461). Comparatively, when only considering the remaining annotators (2,4,6,7 and 8) we find no significant pattern of correlation in this way. These patterns of correlations reiterate an earlier observation of significant ROC-AUC scores for annotators 1,3 and 5, and indicates such patterns between annotators might be a strong source of potential correlated activity that needs to be considered when examining relationships between group-level (combined across annotators) manual annotations and EEG-based predictions [1]. In effect some of these could be considered 'good' annotators.

Importantly, such results indicate (insofar as we can measure) that similar images are arousing similar responses across annotators, it's just that perhaps annotators 4,7 and 8 (and less so 2 and 6) may not be accurately reporting these in a sense we can effectively measure. We find further evidence of this examining examining Spearman's rho correlation of the averaged manual annotation scores (for annotators 1,3 and 5) with the averaged EEG-prediction scores for annotators 2,4,6,7 and 8 (Spearman's rho = .095, α=.039, N=461). Such relationships can be further teased out examining Figure 4 and Figure 3. Taken in tandem these results (from PCA and examining correlative based measures) support the conclusion similar images - at least in part - captured annotator's attention in a similar way although more complex underlying relations do seem to be present.

Although we could theoretically use the annotations provided in phase III as potential inputs in a machine learning scheme to learn models directly from - and apply to - phase I data, this presents a problem as there is high variability across user's annotation counts for the different possible response types (e.g. interesting (target) vs non-interesting (non-target)). This can be seen in Table 1 ('# Targets'). An implication of this is that there can be too few training examples in many instances to adequately train and evaluate a machine learning model. Moreover, the fact a subset of annotators annotate a larger number of images of 'interest' is likely indicating too that annotators are applying different thresholds in how they decide which images to report as having caught their attention (i.e. 'of interest'). Analysis in the presence of such an unequal distribution of label counts across participants could likely be improved by using a likert scale and normalising scores as one method to improve comparability i.e. the annotator rates from 1 to 10 how interesting they found the image and we recalibrate the scores afterwards. Inherent too in this approach, however, is the fact that annotators may equally fail in providing consistent labelling judgements over time. We find evidence of this in our labelling (phase III) task, that is the distribution of target labels made by annotators tend to be made in a way during the annotation task that does not fit with an expected uniformally-tending distribu-

[1] As we are using non-parametric (rank based) statistics in our analysis, there is no difference between averaging a subset of EEG-prediction scores or taking their sum i.e. the respective underlying ranking remains the same for the test. In instances where different N are present (e.g. different consensus N on each image as part of masked EEG-prediction score averaging) we use averaging as here it does make a difference.

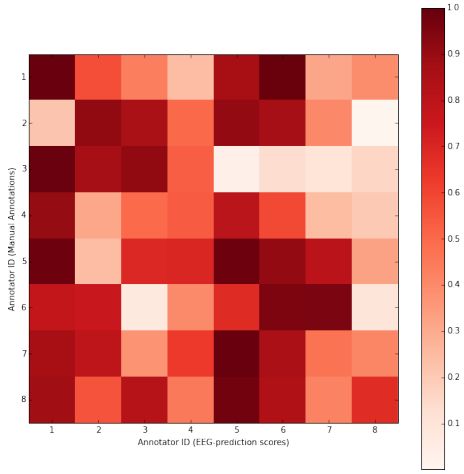

Fig. 3: Mann-Whitney U (p-values) evaluating the relationship between manual and EEG-prediction scores across participants, that is we are comparing distributions of EEG-prediction scores (for annotator on x axis) using annotator labels (y axis) for target v non-targets. P-values are inverted (i.e. 1 - p-value). That is higher values (more red) indicate more significant relationships.

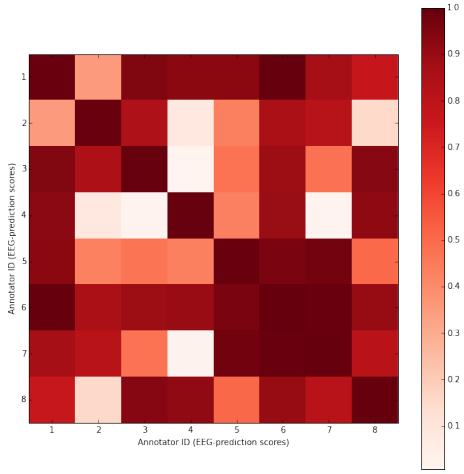

Fig. 4: Spearman-r correlation p-values between participant's for EEG-based annotations (i.e. correlation of EEG-prediction scores). P-values are inverted (i.e. 1 - p-value)

tion across time. As the presentation order of images to be annotated in phase III is randomised we should expect judgements to be made in such a way that labels of a similar type should not be clustered together in time any more than by chance. It would seem annotators are not annotating consistently over time, however, indicating a fundamental problem in how we might naively interpret these labels as being a reliable ground truth. In effect it would seem there might be variation in how decision thresholds are maintained by participants during the task. Using a bootstrapping process with a chi-square statistic examining the expected distribution of target labels during the labelling task we find for annotators 4,5,6,8 this effect of non-uniform distribution is present ($\alpha < .05$). We involve a bootstrapping process as some assumptions for the chi square statistic are violated due to low target counts for some annotators. Comparing how much individual annotators display this effect with respect to each is compromised by the fact of these low target counts meaning in instances with lower target counts and our bootstrapping process, we are increasingly likely to make a type II error.

Table 2: * indicates significant results @ $\alpha < .05$ (via bootstrap resampling randomization test). Consensus N Threshold: The number of annotators that must at minimum agree on an image as a target for it to be included by the threshold as part of the target distribution. Non-Masked MW-U: P-value (Mann-Whitney U test) testing for differences between average EEG-prediction scores (across annotators) for each image i.e. target distribution contains average EEG-prediction scores for all annotators on images where consensus >N. Similarly, non-target distribution comprises average EEG-prediction scores over all images where consensus N = 0. Masked MW-U: P-value (Mann-Whitney U test) as before except only annotators who select the image as a target are included in generating average EEG-prediction scores. Inverted MW-U: P-value (Mann-Whitney U test) except only using averages of EEG-prediction scores for annotators who do not select image as a target.

Consensus N Threshold	Non-Masked MW-U	Masked MW-U	Inverted Mask MW-U	# Pooled Target Counts	# Incremental Target Counts
1	0.063	*0.004	0.194	212	122
2	0.04	*0.007	0.168	90	53
3	0.024	*0.002	0.16	37	22
4	*0.0003	*0.0002	*0.008	15	7
5	0.037	0.059	*0.009	8	5
6	0.242	0.162	0.225	3	2
7	0.056	0.064	*0.043	1	1
8	n/a	n/a	n/a	n/a	n/a

In Figure 2 we can see there becomes relatively fewer images on which there is shared consensus as the consensus count increases across annotators. We define consensus here as the number of annotators that indicate the image was a

target image. Here we sought to investigate whether manual annotations for images with high consensus as being targets across annotators might be indicating those who label such images as non-interesting might be doing so from a neurally informed perspective 'wrongly'. In order to investigate this we use a similar strategy as described above except we examine Mann-Whitney U p-values after applying masks (that is including or excluding certain annotators per image) during the averaging process for EEG-based annotations scores to be examined on a group level. In Table 2 we show consensus counts (# Incremental Target Counts) for each threshold and similarly the number of annotators in agreement for increasing consensus N thresholds using all annotators (# Pooled Target Counts). Important to remember in interpreting results here is that there are uneven contributions from annotators thus averages derived from lower consensus counts tend to be increasingly dominated by annotators with higher target counts from our investigation.

As consensus increases we find fewer sample points available and for this reason in our analysis here (Table 2) we pool inputs to our test (Mann-Whitney U) across incremental thresholds of increasing N (e.g. all examples above or equal to a consensus of 3). For each target image (<consensus N) we calculate an average of EEG-prediction scores of only those annotators who labelled the image as a target (masked), the average prediction scores of those annotators who did not select the image as a target (inverted mask), and an average of all annotator prediction scores. Our comparison distribution (same process) in all cases were images where no annotator selected a target (N=249).

Here a clear relationship emerges in examining pooled targets, masked pooled targets and inverted masked pooled targets. We can see as consensus (N) increases, greater (significantly detectable) effects emerge that shared annotator consensus on an image indicates those who labelled it oppositely (non-target), that their EEG-prediction scores indicate otherwise. Where consensus $N = 4$ and $N = 5$ we can see from the inverted masked analysis high consensus on an image results in greater prediction scores (i.e. target) for annotators who did not actually annotate the image as a target. As discussed, as the consensus N increases we have fewer (target) examples which in turn increases our likelihood of a type II error explaining weaker/non-existent significance at higher N. Similarly for low consensus N we do not see this relationship present.

5 Perspectives on the Experiment and Suggestions

In our experiment with 8 researchers, it seems only 3 of them were careful enough to translate recognition of an object of interest into an actual significantly detectable annotations (as per ROC-AUC). It is our conjecture that in a crowd-sourced environment where annotators are being paid to annotate as many images or items as possible, then the annotator effort and accuracy is likely to be low. This suggests the need for significant annotator performance assessment before relying on the coverage of the annotations and in this respect EEG could be a useful tool in understanding the user's annotation process and

how it relates to neurophysiological correlates of perception and attention. As EEG can be used to index other known neurophysiological phenomena linked to aspects of decision making there is potential too for applications that incorporate error-related responses during the annotation task, that is capturing whether the annotator might have relabelled an earlier annotation as erroneous from examining EEG signals around the time of the response [12].

Another suggestion is the potential for the application of EEG-based annotation that does not require the physical motion of an annotator to press a keyboard or mouse. The observation of the EEG signals is that it is feasible to annotate images at a rate of 5 per second using an EEG-based approach. In experimentation, we have found that such an annotation task can last for approximately 30 minutes (with rest periods) before the average annotator gets too fatigued. However, the application of EEG sensing can easily be done incorrectly such as failing to acquire clean signals or failing to remove confounding/detrimental artefacts from the recording in preprocessing stages such as applying ICA (Independent Component Analysis) or band-pass filtering. These confounding sources of noise due to eye blinks for instance can provide discriminative information in a prediction model but are not of neural origin so are not considered in our work.

These suggestions have the potential to de-risk the annotation process and potentially to significantly improve the speed of annotation and the coverage of the annotations. We suggest that this has implications for the use of crowdsourced annotations and we have made a number of suggestions about the potential for enhanced annotators and EEG-based annotation. Although our preliminary results here are specifically tied to a single dataset and task type, we use these results as a basis to further advocate and discuss the use of neural measures in multimedia research and particularly in tasks involving annotation.

6 Acknowledgements

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

References

1. Ambati, V.: Active learning and crowdsourcing for machine translation in low resource scenarios (2012), aAI3528171
2. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Information Retrieval 11(2), 77–107 (2008)
3. Gerson, A.D., Parra, L.C., Sajda, P.: Cortically coupled computer vision for rapid image search. IEEE Trans Neural Syst Rehabil Eng 14(2), 174–179 (Jun 2006)
4. Healy, G., Smeaton, A.: Eye fixation related potentials in a target search task. In: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. pp. 4203–4206 (Aug 2011)

5. Healy, G., Gurrin, C., Smeaton, A.F.: Lifelogging and eeg: utilising neural signals for sorting lifelog image data. Quantified Self Europe Conference, 10-11 May 2014, Amsterdam, Netherlands. (2014)
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding pp. 675–678 (2014), http://doi.acm.org/10.1145/2647868.2654889
7. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2(1), 1–19 (2006)
8. Mohedano, E., Healy, G., McGuinness, K., Giró-i Nieto, X., O'Connor, N., Smeaton, A.: Improving object segmentation by using eeg signals and rapid serial visual presentation. Multimedia Tools and Applications pp. 1–23 (2015), http://dx.doi.org/10.1007/s11042-015-2805-0
9. Noronha, J., Hysen, E., Zhang, H., Gajos, K.Z.: Platemate: Crowdsourcing nutritional analysis from food photographs pp. 1–12 (2011), http://doi.acm.org/10.1145/2047196.2047198
10. Polich, J.: Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 118(10), 2128–2148 (Oct 2007)
11. Shenoy, P., Tan, D.: Human-aided computing: Utilizing implicit human processing to classify images. In: CHI 2008 Conference on Human Factors in Computing Systems (2008)
12. Spuler, M., Niethammer, C.: Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity. Front Hum Neurosci 9, 155 (2015)
13. Welinder, P., Perona, P.: Online crowdsourcing: Rating annotators and obtaining cost-effective labels pp. 25–32 (June 2010)