Improving Decision-making based on Visual Perception via a Collaborative Brain-Computer Interface

Riccardo Poli*, Caterina Cinel*, Francisco Sepulveda* and Adrian Stoica†
*Brain-Computer Interfaces Lab, School of Computer Science and Electronic Engineering, University of Essex
†NASA Jet Propulsion Laboratory

***Abstract*—In the presence of complex stimuli, in the absence of sufficient time to complete the visual parsing of a scene, or when attention is divided, an observer can only take in a subset of the features of a scene, potentially leading to poor decisions. In this paper we look at the possibility of integrating the percepts from multiple non-communicating observers as a means of achieving better joint perception and better decision making. Our approach involves the combination of brain-computer interface (BCI) technology with human behavioural responses.**

To test our ideas in controlled conditions, we asked observers to perform a simple visual matching task involving the rapid sequential presentation of pairs of visual patterns and the subsequent decision as whether the two patterns in a pair were the same or different. Visual stimuli were presented for insufficient time for the observers to be certain of the decision. The degree of difficulty of the task also depended on the number of matching features between the two patterns. The higher the number, the more difficult the task.

We recorded the response times of observers as well as a neural feature which predicts incorrect decisions and, thus, indirectly indicates the confidence of the decisions made by the observers. We then built a composite neuro-behavioural feature which optimally combines these behavioural and neural measures.

For group decisions, we tested the use of a majority rule and three further decision rules which weigh the decisions of each observer based on response times and our neural and neuro-behavioural features. Results indicate that the integration of behavioural responses and neural features can significantly improve accuracy when compared with individual performance. Also, within groups of each size, decision rules based on such features outperform the majority rule.

I. Introduction

A. Limits of Visual Perception

The human visual system is far superior to any automated computer system in the processing and interpretation of visual scenes in ordinary conditions. However, in the presence of complex scenes, in the absence of sufficient time to complete the visual parsing, or when attention is divided, the human visual system is far from perfect [20], [5], [18], [9]. In these conditions, observers can only typically attend a subset of the features of the scene, thus affecting their ability of accurately assessing situations, which may result in suboptimal decisions.

These limitations can partly be overcome if multiple individuals are involved in the assessment and decision process.

B. Decision Making in Groups

Years of research on decision making have shown how group decisions can be superior compared to individual decisions in many different contexts (see, for example, [8], [13], [15], [14]), including settings where individuals are involved in visuals tasks [26]. However, there are circumstances in which group decision-making can be disadvantageous [12], [4]. Also, even when there is a group advantage, that advantage does not always necessarily increase monotonically with the number of group members, the optimal group-size depending on the task at hand [16]. This can be caused by, for example, difficulties in coordination and interaction between group members, reduced member effort within a group, strong leadership, group judgement biases, and so on [13], [26], [14].

Many of these phenomena are mediated by communication and feedback, whereby members of a group share information and get to know other members' decisions [28]. It is not necessarily the case that the more communication and feedback, the better. A recent study [1], for example, has shown that in some conditions combining information from different freely-communicating individuals is an advantage, while in others, e.g., when there are time constrains or if leadership prevails, communication is an obstacle to optimal decision-making.

The situation is similar for computer-assisted decision-support systems, which can either help or hinder decision-making and situation awareness [3], [2], [30], [7].

C. Neural Correlates of Decision Making and Collaborative Brain-Computer Interfaces

The analysis of brain activity through neuro-imaging and other techniques can reveal important information about the decision-making process. For example, electroencephalography (EEG) can provide information about levels of confidence of decision making [25], attention selection processes [11], conscious or unconscious error detection [19], timing of decisions [29] — all relevant factors of efficient decision making. Also, the neural correlates of an individual decision can be detected as early as about 800 ms before an explicit response is given, as shown for example by [29]. We should also note that it has been known for a long time that other (behavioural) measurements, such as the response times, are influenced, and thus can reveal, the confidence in a decision [17].

Given these psychophysiology findings, it would seem reasonable to attempt to exploit this information to improve decision making (e.g., decisions based on neural activity can potentially be faster or better). However, EEG data are too noisy to use neural correlates on their own to reliably provide information on (or aid) single decisions (all the previously mentioned reports base their findings on averaging the signals resulting from a large number of repetitions of each decision).

Nonetheless, it is plausible to think that collective decisions could be based, or partly based, on the integration of the brain activity of the members of a group. In fact, bypassing overt interaction and communication between group members might help overcome some of the drawbacks of group decision-making, previously discussed, while still preserving a key benefit of group interactions: that individuals who are not very confident in their decision will influence the group's decision less and *vice versa*. Group decision-making supported by the integration of neural activity would be particularly suitable – but not limited – to circumstances where decisions are based on a rapid and accurate assessment of the environment and where fast reactions are needed.

The possibility of aggregating the brain activity of members of a group to reach optimal decisions has been recently explored by [10], who integrated (offline) the EEG activity of up to 20 individuals engaged in a simple perceptual decision-making task (i.e., discriminating between rapidly presented pictures of cars and faces). It was found that combining neural activity across brains resulted in decisions as accurate as those obtained by aggregating the observers' opinions. Also, group decisions could be predicted not only by the neural activity related to the decision processes, but also by the neural activity correlated to early perceptual processing, thus showing that "multi–brain" decisions can be taken faster than decisions based on overt communication.

The idea of multi–brain collaborative decision as proposed by [10] has recently been applied to Brain Computer Interface (BCI) research. [31] have compared the performance of single and collaborative BCI in a task of movement planning. In the experiment described, through directly extracting information from the posterior parietal cortex and bypassing the motor related procedures, the BCI system could accelerate a motor response by using an artificial limb. The results show that there can be an advantage on the overall performance when brain activity of a group of individuals is integrated, compared to single performance, and the larger the group the better the overall performance (groups of up to 20 people were tested in both [10] and [31]). Similarly [32] have proposed an online collaborative BCI that, using visual evoked potentials (VEP), can accelerate human response to visual stimuli.

Multi-brain aggregation not only can facilitate rapid analysis of the environment and decision making, but can also assess characteristics such as group emotions, as shown in [27]. There an experiment was described in which a group's emotional index was obtained by aggregating EEG and electromyographic signals from two individuals who were observing emotion-triggering images.

Very recently [21] we have started studying the potential of a collaborative approach to BCI, too. In particular, we developed collaborative versions of our analogue BCIs for real-time 2–D pointer control developed in previous work [6], [23], [24] where the pointer movement is controlled via the integration of the brain activity of two users. The analysis of performance with three subjects indicated that our best collaborative BCI produces trajectories that are statistically significantly superior to those obtained in the single-user case.

D. Contributions of the Present Study

In the present study we examine the possibility of using neural and behavioural features to improve group decisions in a visual-matching task, where images were presented to observers in taxing perceptual conditions (namely, high perceptual load and high speed of stimulus presentation). As previously discussed, in these cases human perception may not only be incomplete but also incorrect or, at the very least, imprecise. By integrating the percept-related neural activity from multiple observers we hoped to achieve more accurate evaluations of such images as the different observers could possibly detect different visual parts and features of an image.

BCI has hitherto implied the use of brain signals from a single user, but, as seen in Sec. I-C, the technology also gives us access to data pertaining to various cognitive processes which has only recently begun to be investigated in multi-user scenarios. In this paper we introduce a hybrid collaborative BCI, that involves the combination of pure brain-computer interface technology with human behavioural responses in a multi-user setting.

II. Methodology

Our collaborative BCI involves the combination of three features: (a) the neural features extracted from the EEG signals of each group member, (b) the decisions made by each member and (c) the response time. As indicated above, response times are indicators of confidence (longer decision times being normally associated with lower confidence). Also, our neural features were specifically designed to represent confidence. So, by weighing each member's decisions using these features, before algorithmically combining decisions, was hoped to provide more accurate group decisions.

To test our ideas in a suitably constrained environment, we used a particularly simple set of visual stimuli, which, however, were presented very briefly thereby making the matching task particularly arduous.

A. Stimuli and Tasks

Participants underwent a sequence of 8 blocks of trials, each block containing 28 trials, for a total of 224 trials. Each trial (see Fig. 1) started with the presentation of a fixation cross in the middle of the screen for 1 second. This time allowed participants to get ready for the presentation of the stimuli and allowed EEG signals to get back to baseline after the response from previous trials. Then observers were presented with a sequence of two displays, each showing a set of shapes. The first set (Set 1) was presented for 83 ms (5 frames of a 60Hz screen) and was immediately followed by a mask for

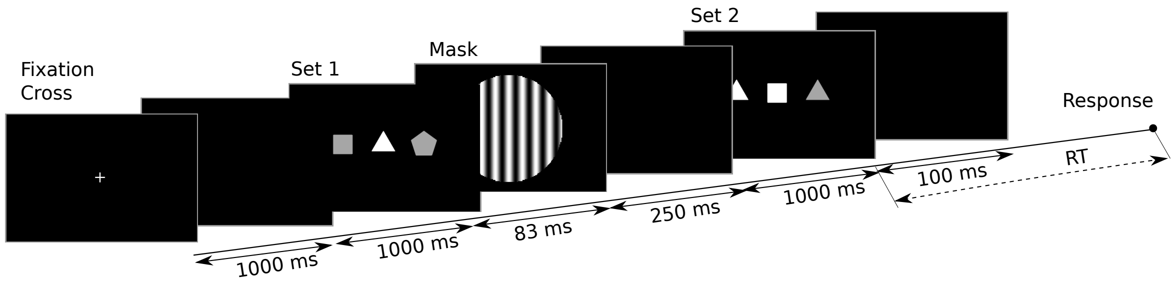

Fig. 1. Stimulus sequence used in our experiments.

250 ms. After a delay of 1 second, the second set of stimuli (Set 2) was shown for 100 ms. Following this, observers had to decide, as quickly as possible, whether or not the two sets were identical. Responses were given with the two mouse buttons (left for "identical", right for "different"), and response times (RTs, expressed in seconds) were recorded.[1] Each set consisted of three shapes (each approximately subtending 1 degree and being approximately 1.8 degrees apart), which could be any combination of a triangle, square and pentagon (see Sets 1 and 2 in Fig. 1). Note that the same shape was allowed within the same set. Each shape was coloured either in pure white (corresponding to normalised RGB (1,1,1)) or light grey (RGB (0.65,0.65,0.65)). Shapes were presented on a black background.

With two shades of grey and 3 possible shapes for each of the three elements in each set, there were a total of $(2 \times 3)^3 = 216$ different possibilities for each set, leading to a $216^2 = 46,656$ possible set combinations. Since, each element of the three stimuli in a set has two features (grey level and shape), we divided the set pairs based on the number of matching features they shared, a number that we called *degree of match* (DoM). If all three stimuli of Set 1 differ in both shape and grey level from the three stimuli in Set 2, we have a DoM of 0; if one element shares a feature (e.g., the same shape), that is a DoM of 1; etc. So, DoM ranges from 0 to 6 (6 corresponding to a perfect match between Set 1 and Set 2).[2]

The combination of the shapes in Set 1 and their grey levels were randomly selected. However, we found that randomly selecting also the features of Set 2 would produce a disproportionate number of sets which had an intermediate DoM, thereby under-representing the cases where a decision is particularly difficult and also the "identical" condition. So, we imposed a constraint that while stimuli would be random, there should be equal proportions of each DoM in each block. Once randomly generated, the sequences of sets were stored, so that identical sequences were used for all subjects.[3]

[1]Participants were given plenty of practice with the task and the response buttons before the experiment. We also placed labels reminding them of which mouse button was to be used for which response at the bottom of the screen.

[2]Note that a feature was "shared" only when it was in the same position in the two sets. Therefore, if, for example, Set 1 showed a triangle in the first position, while Set 2 showed a triangle in the second or third position, but not in the first position, that was not a shared feature.

[3]There are two reasons for this. Firstly, this ensures that all subjects underwent exactly the same experiment, which should increase repeatability and reproducibility. Secondly, as we will explain later, this allowed us to test offline the benefits of combining the decisions of multiple users when presented with the identical displays.

B. EEG Signals Acquisition and Transformation

We gathered data from 7 participants with normal or corrected-to-normal vision (average age 33.3, SD 10.3; 5 female, 5 right handed). Participants were seated comfortably at about 80 cm from an LCD screen. EEG data were collected from 64 electrode sites using a BioSemi ActiveTwo EEG system. The EEG channels were referenced to the mean of the electrodes placed on either earlobe. The data were initially sampled at 2048 Hz, the band-pass filtered between 0.15 and 16 Hz and down-sampled to a final sampling rate of 32 Hz. Briefing and preparation of participants, checking and correcting the impedances of the electrodes and task familiarisation took approximately 30 minutes.

The EEG data were segmented into epochs for the purpose of extracting our neural feature. Normally in event-related potentials (ERPs) epochs start with the stimulus (they are "stimulus-locked" in electrophysiology jargon) and last for a certain time. However, here we were also interested in the neural processing that immediately precede a participant's response. This is best captured by using a "response locked" approach, where an epoch is extracted which lasts for a certain amount of time and ends with the response given by the user.

So, we decided to acquire two epochs of data: one lasting 1000 ms and centred on the onset of Set 2 (we wanted to capture data preceding this stimulus since they might reflect the degree of attention devoted to the stimulus itself) and one lasting also 1000 ms and preceding the response (i.e., the response time ending the epoch). The Eigenbrain transform was applied to each epoch [22]. This is similar to principal component analysis, but its basis functions tend to represent larger-scale potentials as illustrated in Fig. 2. So, each epoch now consisted of a set of 32 time samples (1000 ms worth of data at a sampling rate of 32 Hz) for each of the 64 Eigenbrain-transform coefficients.

C. Novel Neural and Neuro-behavioural Correlates of Confidence in Decision-making

We wanted to identify a neural feature which would represent the degree of certainty with which a user takes decisions. However, ground truth information on confidence is not directly available. One can ask a subject to tell his or her degree of confidence in a decision, but it isn't clear how objective subjects would be in doing so. So, here we concentrated on trying to find a more objective surrogate for confidence. In particular, we focused on trying to characterise the cases where

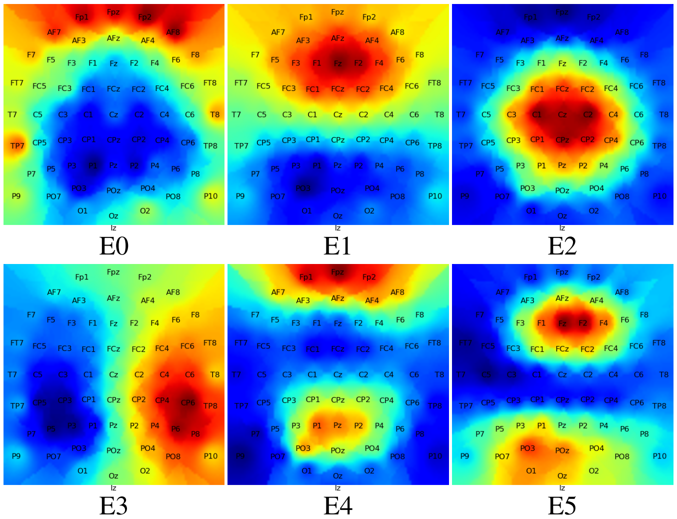

Fig. 2. Scalp-map representation of the six lowest-order (eigen) vectors of the Eigenbrain transform used for pre-processing EEG data in this work.

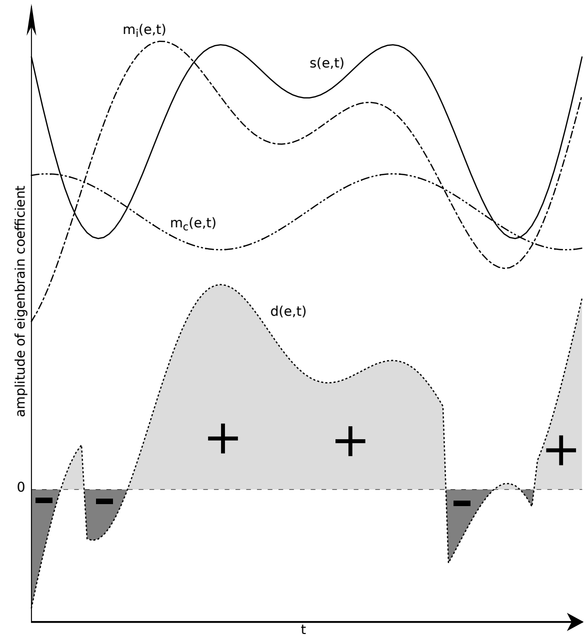

Fig. 3. The rescaling and sign changes which we perform on an epoch $s(e,t)$ to compare it with the prototype for the correct $m_c(e,t)$ and incorrect $m_i(e,t)$ epochs.

the response given by a user was correct *vs* the cases where the response was incorrect.

In a rational observer, we can safely assume that incorrect responses are incorrect because the perceptual processes leading to the decision did not provide all the necessary information to take the correct decision. It thus stands to reason that in these conditions the confidence with which an observer would take a decision would be low for most of the "incorrect" trials. On the contrary, the confidence with which an observer would take a decision would be higher for most of the "correct" trials.[4] Thus, if we could find predictors of whether the decisions made by a subject would be correct or incorrect, we would essentially amount to finding predictors of the degree of certainty of the subject in making such decisions.

We started from forming a prototype response for each of these cases. So, we divided up our (Eigenbrain-transformed) epochs into those were a correct response was given and those were it wasn't, and computed the medians of across trials (on a time sample by time sample and Eigenbrain coefficient by Eigenbrain coefficient basis) for the two categories: "correct" and "incorrect". We will call these $m_c(e,t)$ and $m_i(e,t)$, respectively, where e represents the e-th Eigenbrain coefficient and t represents the t-th sample. If $s(e,t)$ is a particular epoch that we want to analyse, we transform it into a *signed-weighted difference* function $d(e,t)$ by scaling the epoch and comparing it to the two prototypes. This is defined as follows:

$$d(e,t) = \text{sign}(m_i(e,t) - m_c(e,t)) \times (s(e,t) - m_c(e,t)) \quad (1)$$

Let us look more closely at this function.

As illustrated in Fig. 3 (where all signals are artificial examples, for clarity of visualisation), the prototypes Eigenbrain coefficient amplitudes for the correct epochs, $m_c(e,t)$, and for the incorrect $m_i(e,t)$ epochs may not be very well separated and, in fact, may cross each other. So, it isn't immediately obvious how one could compare a Eigenbrain-transformed epoch, $s(e,t)$, with such prototypes and determine to which of the two the epoch resembles more. If we look at the sign-weighted difference function, $d(e,t)$, we see that it is 0 whenever $s(e,t)$ is identical to the prototype for the correct epochs, $m_c(e,t)$, i.e., when the line $s(e,t)$ crosses the line $m_c(e,t)$. We also see that the sign-weighted difference is positive (represented by the lighter shaded areas in the figure) when $s(e,t)$ is closer to the prototype for the incorrect epochs, $m_i(e,t)$, than to $m_c(e,t)$. Also, the sign-weighted difference is negative (this is represented by the darker shaded areas in the figure) when $m_c(e,t)$ falls in between $s(e,t)$ and $m_i(e,t)$, that is, when $s(e,t)$ is more further away from the prototype for the incorrect trials than the prototype for the correct trials. So, in essence $d(e,t)$ tells us, coefficient by coefficient and sample by sample, how much the value of the coefficient is more towards one prototype than the other.

Since we are interested in looking at the similarity between the potentials and the prototypes at all samples in an epoch and for all Eigenbrain components, we average the sign-weighted difference $d(e,t)$, obtaining the following (scalar) *neural feature*

$$nf = E[d(e,t)] \quad (2)$$

where the expectation operation, $E[.]$, is applied to both e and t. Like for $d(e,t)$, large positive values of *nf* indicate more similarity of an event related potential (ERP) with the prototype for incorrect responses than for correct ones. On the contrary, small positive values and, even more, negative values indicate more similarity of the ERP with the prototype

[4] Of course, when one is not confident as to what he or she has seen, random guessing may be a significant element in the decision. When an observer guesses, it is possible that he or she will give the correct response just by shear luck. So, a fraction of the trials where a correct response is recorded may be characterised by low confidence in the decision. However, if the proportion of correct decisions is sufficiently high (as in our experiments), in the majority of "correct" trials the observer's confidence will be significantly higher than for "incorrect" trials.

for correct responses than for incorrect ones.

The values of *nf* were normalised on a subject by subject basis by subtracting their median and dividing by the median of the absolute deviations from the data's median (also known as MAD), thereby obtaining the values nf_n. We also normalised the response times, *RT*, by dividing them by their median, thus obtaining the values RT_n. We used these statistics for normalisation as they are much more robust to outliers than the mean and standard deviation.

Finally, we defined a composite feature, *RTnf*, as the optimal linear combination of RT_n and nf_n. That is:

$$RTnf = a \times RT_n + b \times nf_n + c \tag{3}$$

where a, b, and c are constants that were optimised by linear regression on a subject by subject basis. For the regression, the values of the dependent variable were $+1$ for trials resulting in an *incorrect* response and -1 otherwise. So, on a trial by trial basis, *RTnf* is a prediction of how likely the response given by a subject will be incorrect, with the higher the value of *RTnf* the higher the probability of error. Indirectly, however, as we discussed above, the variables nf_n and *RTnf* provide a trial-by-trial measurement of a subject *uncertainty* in making a decision.[5]

D. Learning Neural and Neuro-Behavioural Features

Our techniques rely on extraction of information and the statistical analysis of the data (ERPs, response times and actual responses given by the subjects). This is a form of machine learning. As such we must ensure that our results are not affected by over-fitting. So, validation of our methods requires splitting our data into a training and a test set. Because the data sets one can acquire in electrophysiology, neural engineering and BCI studies are always relatively small compared to other domains, we adopted, as is customary, an *8-fold cross-validation* approach. In each fold we had a training set of 196 trials and a test set of 28 trials and we estimated the prototypes, the normalisation statistics and the linear regression coefficients used in the features only using the training set. We then reused the same values to estimate our three features RT_n, nf_n and *RTnf* for the trials in the test set.

E. Using Behavioural, Neural and Neuro-behavioural Features in Decision-making

As we indicated above, our objective was to combine the behavioural and neural features from multiple users to see under what conditions their decisions in a perceptual task would be most improved with respect to those taken by a single observer in conditions of complete absence of communication or any other form of social influence. To achieve this, we decided to compare the standard majority rule against rules where the confidence of the observers (as assessed by our three features RT_n, nf_n and *RTnf*) is used to weigh their decisions. In the case of majority, each observer's decision is a vote and the final decision is based on straight majority for teams with an odd number of members and majority followed by the flipping of an unbiased coin in the case of ties for teams with an even number of members.[6] In all the other cases we simply considered each decision as being worth w_i, w_i being the weight associated with observer i (a value that is computed a trial by trial basis), and we take the following joint decision:

$$\textit{decision} = \begin{cases} \text{yes} & \text{if } \sum_{i\in\mathcal{Y}} w_i > \sum_{i\in\mathcal{N}} w_i, \\ \text{no} & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{Y}$ and $\mathcal{N}$ represent the sets of all observers who decided "yes" and "no", respectively.

The weights w_i were chosen as follows. When RT_n was used as a measure of confidence we set

$$w_i = \frac{1}{RT_n}, \tag{5}$$

when nf_n was used we set

$$w_i = \frac{1}{\max(nf_n + 5, 0.5)}, \tag{6}$$

and when *RTnf* was used we set

$$w_i = \frac{1}{\max(RTnf_n + 5, 0.5)}. \tag{7}$$

Since nf_n and *RTnf* may take arbitrarily large positive or negative values (e.g., in the presence of artifacts cause by eye blinks, swallowing, etc.), the transformation $x \rightarrow \max(x + 5, 0.5)$ was used to prevent the denominators in Equations (6) and (7) from ever becoming negative. The value 0.5 was used instead of just 0.0 to prevent any w_i from becoming too big, which would potentially lead to one observer's vote counting many times more than the rest of the group together. This wasn't necessary for the *RT* feature, because response times are always positive and can never be too short (remember: they are normalised, so their median is 1).

III. Results

A. Behavioural, Neural and Neuro-behavioural Features

To investigate the relationship between correct/incorrect responses and the confidence with which decisions were taken, we studied the distributions of normalised response times, as well as the neural and neuro-behavioural features defined in the previous section. We binned the data (obtained via cross validation) on the basis of two criteria: (a) the degree of match of the stimuli presented in a trial (as an indicator of the objective difficulty of the task of discriminating them), and (b) whether the corresponding decision taken by a subject was correct or incorrect.

Table I reports the medians (across all subjects) of the behavioural feature RT_n, the neural feature nf_n, and the neuro-behavioural feature *RTnf* as a function of the DoM of the

[5] In the experiments, we actually put this interpretation to the test by verifying whether, instead, these variables could be used as measures of *certainty*. Results of joint decisions were exceptionally bad when we adopted this criterion, confirming our original line of reasoning.

[6] To reduce the noise in our performance estimates we didn't actually flip a coin. Instead, we used the expected value of the outcome of the decision. That is, when counting the number of correct decisions we added 0.5 to the count for every decision where there wasn't a majority since such decision would turn up to be correct in exactly 50% of the cases.

TABLE I

MEDIANS (ACROSS ALL SUBJECTS) OF BEHAVIOURAL, NEURAL AND NEURO-BEHAVIOURAL FEATURES AS A FUNCTION OF THE DEGREE OF MATCH OF THE STIMULI USED IN A TRIAL.

DoM	RT_n	nf_n	*RTnf*
0	0.792227825325	-0.0510559687187	-0.836577642217
1	0.809819204299	-0.1309891896820	-0.880723356710
2	0.819614990349	-0.1855942629640	-0.872865696435
3	0.860335354760	0.1432751068230	-0.827987399221
4	0.928805366451	0.0821520018159	-0.776855148120
5	0.999058005841	0.2849991108550	-0.757815272268
6	1.092691431290	0.6328967131570	-0.669984517787

TABLE II

MEDIANS (ACROSS ALL SUBJECTS) OF BEHAVIOURAL, NEURAL AND NEURO-BEHAVIOURAL FEATURES AS A FUNCTION OF WHETHER THE USER'S DECISION WAS CORRECT OR INCORRECT IN A TRIAL.

Decision	RT_n	nf_n	*RTnf*
correct	0.851246791855	0.0445507448855	-0.825989723237
incorrect	1.183397160150	0.5753178988980	-0.623682337966

stimuli used in a trial. Box plots representing the distributions of the features for different values of DoM are reported in Fig. 4(top). Overall we see that the medians of the features tend to monotonically increase as DoM increases, indicating that they correlate with the difficulty of the task.

Table II reports the medians of the same features as a function of whether the users' decisions were correct or incorrect. The corresponding box plots are shown in Fig. 4(bottom). As one can see from these, the medians are much higher for the incorrect decisions than for the correct ones for all the features used. These differences resulted to be highly statistically significant when we applied the one-tailed Kolmogorov-Smirnov test to the data ($p < 10^{-5}$ for all comparisons). This indicates that observers found the decisions characterised by higher values of our features more difficult than those characterised by lower values.

Overall these results support the hypothesis that our features represent reasonable indicators of decision confidence.

B. Group Decisions

We compared the performance of single observer decisions with group decisions within groups of increasing size. All possible memberships of the groups were tested. Since we had 7 observers, we had 7 "groups" of size 1, 21 groups of 2 observers, 35 groups of 3 observers, 35 groups of 4 observers, 21 groups of 5 observers, 7 groups of 6 observers, and one group of 7 observers.

For each group we computed the number of errors made by the group when using the four different methods of making decisions studied in the paper (i.e., based on majority and our three features RT_n, nf_n and *RTnf*). For each group size we then computed the mean number of errors made with each method.

In Fig. 5, we report the average percentage of errors as a function of group size for the four methods for group decisions tested in the paper. As one can see, in all methods studied, group decisions were superior to the decisions of single observers, suggesting that integration of perceptual information across non-communicating observers is possible

TABLE III

p-VALUES RETURNED BY THE ONE-TAILED TWO-SAMPLE KOLMOGOROV-SMIRNOV TEST WHEN COMPARING THE PERFORMANCE OF SINGLE OBSERVERS AGAINST THE PERFORMANCE OF GROUPS OF DIFFERENT SIZES AND ADOPTING DIFFERENT DECISION METHODS. VALUES BELOW 0.01 ARE IN BOLD FACE.

Group Size	Majority	RT_n	nf_n	*RTnf*
2	0.807118	*0.092462*	0.311403	*0.074818*
3	**0.000259**	**0.000456**	**0.000262**	**0.000264**
4	**0.000189**	**0.000189**	**0.000189**	**0.000189**
5	**0.000446**	**0.000446**	**0.000446**	**0.000446**
6	**0.005841**	**0.003362**	**0.005841**	**0.000912**
7	0.173774	0.173774	0.173774	0.173774

and beneficial. Also, we see that the straight majority is either on par or outperformed by the other three methods. This is particularly evident with groups with an even number of members where the coin-tossing decisions taken in the presence of ties imply that performance is the same as that of groups with one fewer member. The figure also shows that of the three other methods, the one based on RT_n and $RTnf_n$ appear to be superior to that based on the purely neural feature nf_n.

To test if these observed differences are statistically significant, we also compared the distributions of errors made. In particular, we compared the error distributions of single observers with the error distributions of groups of increasing size (for the four methods of joint decision tested) and the error distributions across the methods within each group size. In the former case, we used the one-tailed two-sample Kolmogorov-Smirnov test (which is a non-parametric test to compare two data samples). In the latter case, since errors are paired in each comparison (by the fact that the two methods being compared were applied to exactly the same groups), we used the more powerful one-tailed Wilcoxon signed-rank test (a non-parametric test used when comparing paired data samples).[7]

Table III reports the p-values returned by the one-tailed two-sample Kolmogorov-Smirnov test when comparing the performance of single observers against the performance of groups of different sizes and adopting different decision methods.[8] This shows that for groups of size 3, 4, 5 and 6, groups decisions are significantly superior to single observers. For groups of size 2, RT_n and *RTnf* are nearly significantly superior to single observers. Note also that our group of size 7 is, unsurprisingly, not superior to single observers, despite its performance being superior to all of the single observers, due to it being a sample of just one data point. It is not unlikely that in both cases the superiority of group decisions could be proven in future larger studies.

We also looked at statistical differences *within* each group size. There we found that several of the small differences

[7]Formally these tests require the data to come from a continuous interval, while the number of errors made by a group is an integer. We, therefore, modified them by adding a very small amount of noise (uniformly distributed in $[-10^{-3}, +10^{-3}]$) to the error data to randomly break ties before applying the tests. To reduce stochastic effects we repeated this process 1,000 times, and computed the average p values recorded.

[8]The values of some of the entries in the table coincide when all or all but one data points are superior to all single observers.

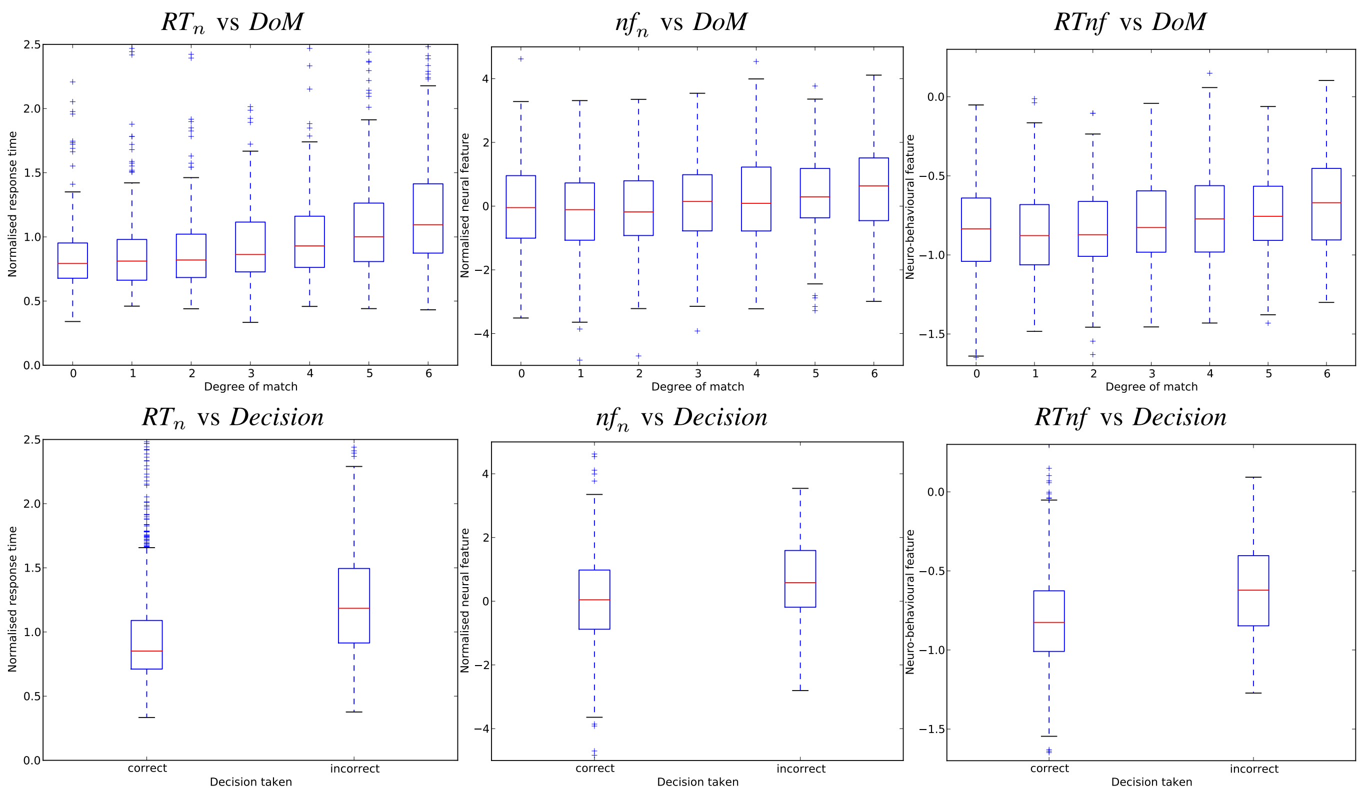

Fig. 4. Box plots representing the distributions of the features for different values of DoM (top) and for different decisions (bottom).

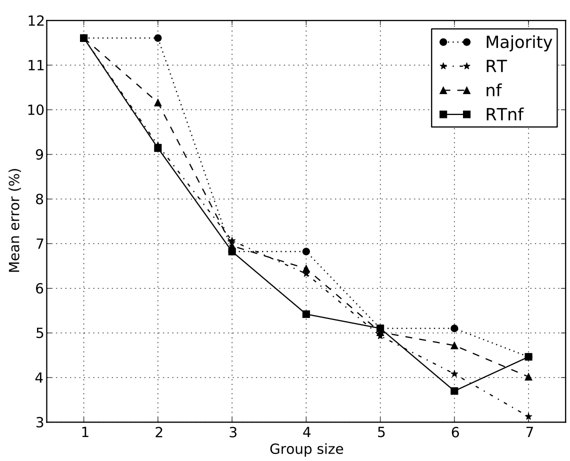

Fig. 5. Average percentage of errors *vs* group size for the four methods for group decisions tested in this paper.

shown in Fig. 5 are not significant. In groups of size 1 (all methods performing the same) and 7 (where we only have one such group), rather obviously, no difference is statistically significant. Also, in groups of size 5 no difference is significant. However, in many other configurations, there are significant differences (all p-values being less than 0.03). For example, in groups of size 2, quite interestingly, RT_n, nf_n and $RTnf$ are all statistically better than majority (also RT_n is better than nf_n). In groups of size 3, majority and $RTnf$ are better than RT_n. In groups of size 4, again, RT_n, nf_n and $RTnf$ are all better than majority. In addition, $RTnf$ is better than RT_n and nf_n. In groups of size 6, RT_n and $RTnf$ are better than majority. Also, $RTnf$ is better than nf_n.

Summarising these results we see that, of the 4 group sizes where there was statistical significance, majority was the worst performing method, being probably superior to only one of the proposed methods (RT_n) and in only one group size (size 3). In all other cases it is either on par or probably worse than our methods. Also, we find that $RTnf$ is never worse than majority in groups of odd size while it beats it in all cases of an even group size. RT_n is a close second and nf_n is third best.

IV. Discussion and Conclusions

The purpose of this study was to investigate whether group decisions based on visual perception would be superior to individual decisions and whether a BCI exploiting neural and behavioural measures of confidence to weight each member decisions, on a decision-by-decision basis would score better than the simple majority rule. Experimental evidence gathered with 7 participants conclusively indicates that both questions can be answered in the positive. On average group decisions showed at least a two-fold reduction in errors with groups of 5 to 7 members irrespective of the decision method used. Also, the results obtained when weighting votes via our neuro-behavioural feature were always either on par or better than those obtained with the majority rule.

Our method offers a number of advantages: a) group decision making can be achieved without communication, thus avoiding the negative effects of human factors discussed previously; b) decisions are based on the integration of the individual decisions with the response times and neural features which correlate with confidence levels, resulting in more accurate decisions; c) the method is applicable and particularly beneficial when applied to groups of an even size, where the standard majority rule would be inapplicable.

In this study observers performed a relatively simple visual matching task, which is nowhere as complex as those carried out in realistic decision-making situations. So, in the future we will need to investigate whether the benefits of our hybrid collaborative BCI approach for group decisions also accrue with more demanding real-world scenarios, with different perceptual modalities (e.g., audio signals) and with more complex decisions. We also will need to see whether it is possible to extend it to perform group decisions where different members are exposed to different sources of information (unlike here, where the were exposed to exactly the same information) thereby providing a form of collective situation awareness.

ACKNOWLEDGEMENTS

The authors would like to thank the UK's Engineering and Physical Sciences Research Council (EPSRC) for financially supporting this research (grant EP/K004638/1).

REFERENCES

[1] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[2] B. B. Baltes, M. W. Dickson, M. P. Sherman, C. C. Bauer, and J. S. LaGanke. Computer-mediated communication and group decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 87(1):156–179, January 2002.

[3] P. Bordia. Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *The Journal of Business Communication*, 34:99–120, 1997.

[4] L. Branson, N. L. Steele, and C.-H. Sung. When two heads are worse than one: Impact of group style and information type on performance evaluation. *Journal of Business and Behavioral Sciences*, 22(1):75–84, 2010.

[5] C. Cinel, G. W. Humphreys, and R. Poli. Cross-modal illusory conjunctions between vision and touch. *Journal of experimental psychology. Human perception and performance*, 28(5):1243–1266, Oct 2002.

[6] L. Citi, R. Poli, C. Cinel, and F. Sepulveda. P300-based BCI mouse with genetically-optimized analogue control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(1):51–61, Feb. 2008.

[7] T. Comes, N. Wijngaards, J. Maule, D. Allen, and F. Schultmann. Scenario reliability assessment to support decision makers in situations of severe uncertainty. In *IEEE Conference on Cognitive Methods in Situation Awareness and Decision Support*, March 2012.

[8] J. H. Davis. Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 80(2):97–125, Mar. 1973.

[9] J. DiVita, R. Obermayer, W. Nugent, and J. M. Linville. Verification of the change blindness phenomenon while managing critical events on a combat information display. *Human Factors*, 46(2):205–218, 2004.

[10] M. P. Eckstein, K. Das, B. T. Pham, M. F. Peterson, C. K. Abbey, J. L. Sy, and B. Giesbrecht. Neural decoding of collective wisdom with multi-brain computing. *NeuroImage*, 59(1):94–108, 2012.

[11] S. A. Hillyard and L. Anllo-Vento. Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America*, 95(3):781–787, February 1998.

[12] D. Kahneman and S. Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 49–81. Cambridge University Press, Cambridge, 2002.

[13] N. L. Kerr, R. J. Maccoun, and G. P. Kramer. Bias in Judgment: Comparing Individuals and Groups. *Psychological Review*, 103(4):687–719, Oct. 1996.

[14] N. L. Kerr and R. S. Tindale. Group Performance and Decision Making. *Annual Review of Psychology*, 55(1):623–655, 2004.

[15] P. Laughlin, B. Bonner, and A. Miner. Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, 88:605–620, 2002.

[16] P. R. Laughlin, E. C. Hatch, J. S. Silver, and L. Boh. Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, 90(4):644–651, 2006.

[17] R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, 1986.

[18] A. Mack and I. Rock. *Inattentional blindness*. MIT Press, Cambridge, MA, 1998.

[19] S. Nieuwenhuis, K. Ridderinkhof, J. Blom, B. G.P.H., and A. Kok. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, 38:752–760, 2001.

[20] J. Palmer. Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):332–350, May 1990.

[21] R. Poli, C. Cinel, A. Matran-Fernandez, F. Sepulveda, and A. Stoica. Towards cooperative brain-computer interfaces for space navigation. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, Santa Monica, CA USA, 19–22 March 2013. (to appear).

[22] R. Poli, L. Citi, M. Salvaris, C. Cinel, and F. Sepulveda. Eigenbrains: the free vibrational modes of the brain as a new representation for EEG. In *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6011–6014, Buenos Aires, Argentina, August 31 – September 4 2010.

[23] R. Poli, L. Citi, F. Sepulveda, and C. Cinel. Analogue evolutionary brain computer interfaces. *IEEE Computational Intelligence Magazine*, 4(4):27–31, November 2009.

[24] M. Salvaris, C. Cinel, L. Citi, and R. Poli. Novel protocols for P300-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, (1):8–17, 2012.

[25] A. Selimbeyoglu, Y. Keskin-Ergen, and T. Demiralp. What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clin Neurophysiol*, 123:1158–1167, 2012.

[26] R. D. Sorkin, C. J. Hays, and R. West. Signal-detection analysis of group decisions. *Psychological Review*, 108(1):183–203, 2001.

[27] A. Stoica. Multimind: Multi-brain signal fusion to exceed the power of a single brain. In A. Stoica, D. Zarzhitsky, G. Howells, C. D. Frowd, K. D. McDonald-Maier, A. T. Erdogan, and T. Arslan, editors, *EST*, pages 94–98. IEEE Computer Society, 2012.

[28] R. S. Tindale, T. Kameda, and V. B. Hinsz. Group decision making. In J. Cooper and M. Hogg, editors, *Sage Handbook of Social Psychology*, pages 381–403. Sage Publications, London, 2003.

[29] A. Tzovara, M. M. Murray, N. Bourdaud, R. Chavarriaga, J. del R. Millán, M. D. Lucia, and M. D. Lucia. The timing of exploratory decision-making revealed by single-trial topographic EEG analyses. *NeuroImage*, pages 1959–1969, 2012.

[30] F. Vachon, D. Lafond, B. R. Vallieres, R. Rousseau, and S. Tremblay. Supporting situation awareness: A tradeoff between benefits and overhead. In *IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pages 284–291, 2011.

[31] Y. Wang and T.-P. Jung. A Collaborative Brain-Computer Interface for Improving Human Performance. *PLoS ONE*, 6(5):e20422+, May 2011.

[32] P. Yuan, Y. Wang, W. Wu, H. Xu, X. Gao, and S. Gao. Study on an online collaborative BCI to accelerate response to visual targets. In *Proceedings of 34nd IEEE EMBS Conference*, 2012.