BSc Thesis

Feasibility study for an EEG-based semantic tagging game

Jeroen Kools [jkools@science.uva.nl]
University of Amsterdam
Bachelor of Artificial Intelligence
Student nr: 0306177
December 23, 2009

Supervisor: dr. Sennay Ghebreab
Intelligent systems lab
Informatics Institute, University of Amsterdam

Abstract

Semantic tagging of images has a broad spectrum of applications, of both scientific and commercial interest. Experience has indicated that designing the interaction with human taggers in the form of a web game can be a very effective way to collect valuable tagged data. So far, implementations of such games have depended on explicit, deliberate input by the users. A novel alternative is presented that doesn't only use signals from mouse and keyboard, but also utilizes brain activity data in the form of electroencephalography (EEG) scans. Brain response data from an earlier experiment [1] is analyzed in an attempt to find and evaluate a measure of similarity between responses of two subjects, suitable for use in a game. From the perspective of the test subjects, depending on the characteristics of the proposed game and the target audience, such a game could spark their interest in a computational approach to cognitive science.

Table of Contents

1. Introduction

This paper explores the feasibility of creating a game-like image processing experiment based on EEG readings. The game should enable a more thorough investigation of the connection between semantic content of a stimulus and features of the brain activity that this stimulus causes. If it can be determined whether or not two subjects are looking at the same image or images from the same category, solely on the basis of evoked brain potentials, that would establish a clear and important relation between brain activity and semantic image content. Accordingly a possible full scale follow-up experiment could implement and use a game based on this principle.

The purpose of the game is twofold. Firstly, a game based on such a method would allow for effective collection of additional brain response data in combination with semantic information. Secondly there is an educational goal. This is because it would enables researchers who also have teaching interests, to reach out and communicate with the test audience to inform them about the more technical, computational approach to cognitive science in an challenging and entertaining environment.

The basic concept of the game would be that two players attempt to guess each other's 'secret' picture from a set of pictures, using only some indicator based on both their brain activity as a measure of performance or correctness.

Obviously, this game I propose will depend on an adequately accurate method to express similarity between two EEG responses before any more details of the experiment or game can be determined. Therefore, the most important question that must be answered regarding this response-to-response similarity measure is: How can analysis of brain activity data supply enough information to reliably judge whether two persons are looking at the same image? Principal component analysis is the preferred method to reveal the maximally relevant factors in the brain responses, and then a correlation or least squares method are the prime candidates to establish similarity between the principal components.

Two related secondary questions can also be identified. First, do the semantic categories of the images correlate with the response brain activity in the test subject? And second, is there a notable difference between brain responses to images with similar semantic content on one hand and images with dissimilar content on the other hand?

These general questions translate into direct objectives to identify, test and evaluate one or more measures of similarity between data representing test subject's brain responses to images. An estimate of the reliability of the methods is an essential part of this. Then I can further elaborate on the game's mechanics and possible refinements. This will however stay a framework; an actual implementation is not within the scope of this paper.

The results of an actual experiment could be relevant for a number of areas. A greater understanding of the processing of image semantics in the brain could be useful for generating and searching content on semantic web, image retrieval, filtering, and user-based content customization. Finally, the game would help creating enthusiasm for the computational approach of cognitive research.

The next section contains an overview of the historical background of the problem and the fields of image processing and computer vision in general. This is followed by a short summary of the most closely related recent research on visual processing and semantic tagging games.

2. Background

After the rise of computer vision research in the 1980’s and growing availability of computing power, ever increasing amounts of data led to a strain of research into content-based image retrieval. Initially, such search methods generally depended on directly accessible, low level features such as colors, shapes and contrast [2, 3] to find a requested image. Gradually, research processed from the most basic and directly available image features to more advanced image statistic features.

In the same period, a number of advances were made in the field of image processing. This approach is primarily inspired by modeling human perception, and a significant amount of research was done on image statistics and the relation between these statistics and image content. Among these inquiries were ones that examined image entropy and orientation [4], contrast [5], and power spectra and the distribution of contrast over direction in natural images [6]. The main disadvantage of this feature-based approach is the semantic gap between the query and the required result.

This semantic gap [7] is the discrepancy in structure and meaning between the query, the user input and the output in the form of a digitalized image. The input is textual and meaningful, but encoded in ambiguous natural language, while images are - to a computer – unintelligible but precise arrays of numbers lacking any kind of description. The task of connecting the concept with the data, providing a means of interpreting semantics and translating data into words, is far from trivial and still unsolved.

This is why the late 1990’s saw a shift of attention to a more meaningful and concept-based approach image retrieval and pattern recognition [8]. Such a semantic route allows for more accurately aimed and more user friendly search. However, the problem remains that the data lack semantic information. There are two basic ways to deal with this. One might try to somehow extract the needed ‘higher’, semantic information from lower level features of the image data itself, or rely on human users tagged the data as having belonging to some category or having certain semantic content.

This last solution can provide accurate semantic information, but also requires a great deal of human effort, which might not seem feasible for tasks with large corpora of images. One method that has shown to be is effective in providing tagged data is using web games as experiments. Semantic tagging web games have some similarities – although the field is rather different - with how distributed computing projects such as SETI@home [9] have allowed solving computationally massive problems by recruiting the resources of the general public, providing information that would otherwise have been hard, expensive or impossible to obtain. But unlike a distributed computing approach, a semantic web game introduces a competitive element and motivates layman users to contribute more data by rewarding correct classifications or tags.

A number of papers by Luis von Ahn [10,11,12] have shown particularly well how the web game paradigm can be used to combine entertainment and science in order to bridge the semantic gap. Because this paper explores the possibilities of such a game in combination with EEG data, I will further elaborate on the theory behind it in following sections.

Another development in the 1990's was an increasing number of methods to measure brain activity. The introduction of fMRI and parallel developments in image processing sparked new interest in EEG measurements. It turned out to be feasible to detect some connections between brain scans and low level cues and semantic categories of stimuli [13]. For example, a test subject looking at a scene with a living, natural subject like a group of birds would have a detectably different brain scan than a subject looking at modern architecture or random noise [14]. It has also recently been shown that by creating a receptive fields model of a measured brain response, it is possible to partially reconstruct the image content of the stimulus [15].

However, combining semantic tagging with EEG responses to visual stimuli has not been done before, and in this paper will try to determine the feasibility of that idea, and describe its possible parameters.

3. Related Work

Luis von Ahn and Laura Dabbish, from Carnegie Mellon University have written several articles about using web games (or 'games with a purpose, www.gwap.com), to collect various kinds of scientific information related to bridging the semantic gap. The games collect data ranging from semantic tags for musical fragments and music videos, to describing words with synonyms, to outlines of objects in a picture. What they all have in common is that the player is coupled with another human with whom they have to compete and cooperate in a fresh and attractively designed, fun environment. By rewarding players for achieving consensus, these games make effective use of the public's desire for entertainment.

A rather different way to determine the semantic content of an image has been explored by Ghebreab, Scholte et al [16]. Using principal component analysis and a variety of statistical methods on low-level cues they identify the parameters of a Weibull distribution, which correlates with both the semantic broad content category of the image and the brain activity of the subject. They describe a model that can reliable determine the image corresponding to a given brain response.

The game approach can quickly collect a large amount of data, but since it does not depend on any statistical analysis of low-level features, it is unable to say anything about new, unseen images. A game where players match images while in the background their brain activity is analyzed and matched with image statistics would be the best of both worlds. If there is a connection in the data that has already been collected between sensor readouts

4. Methods

4.1 Measures of similarity

This section will outline the specifications of the used EEG data, the experiment they were collected in, and several methods and measures used to analyze the dataset. After that I describe a possible design for a game that would use meet the requirements regarding the type of data that is gathered and uses this measure of EEG similarity.

The data was gathered for an earlier experiment [14]. In that experiment, a group of 33 subjects were each briefly exposed to a set of 800 same-sized color images. This set is divided in clear and broad semantic categories, containing 400 scenes with animals and 400 natural and manmade scenes of non-animals. Each subject was individually exposed to every image for a very brief time. By limiting exposure to 100 milliseconds, only lower-level, subconscious response will be recorded by a skull cap of 64 EEG sensors measures local brain potential more than twice every millisecond. This results in a 218x64 matrix for every trial (see Figure 1), and 33 different 800x218x64 matrices in total.

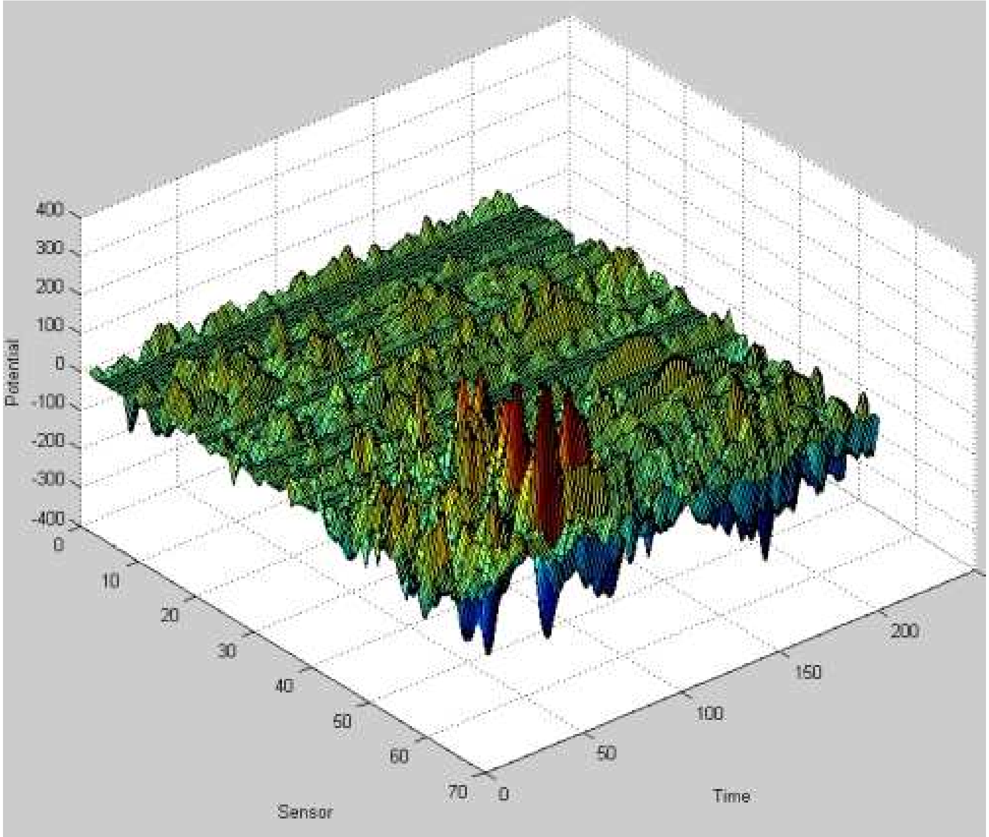

Figure 1: A typical response matrix. Visual inspection shows that there is clear, strong response especially in the first 100 measurements. The highest potentials are recorded by the sensors between locations 50 and 60, especially sensor 58.

To deal with the large and noisy amount of data in this multivariate dataset, principal component analysis (PCA) is applied to each brain response and to the averaged brain response to each image, in order to find the factors that best explain the variance and in the data. PCA would transform a person's 218x64x800 response matrix into a matrix of 800x800 components, with each column a component ordered by importance. Similar brain responses can be expected to have similar principal components.

Therefore the difference and correlation between principal components will be used in several ways in order to establish a similarity measure.

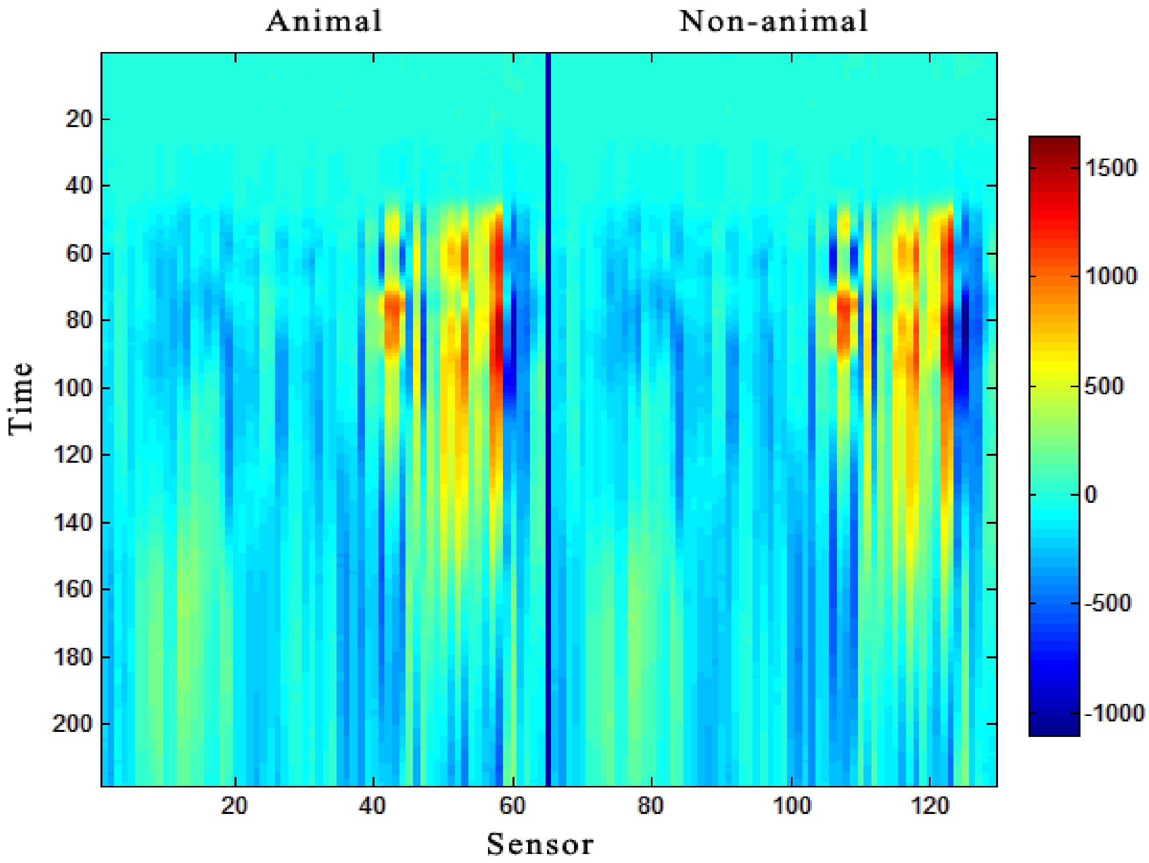

Figure 2: Cross-subject cross-category average response patterns seem to be highly similar on visual inspection. PCA brings out the most relevant differences.

If this 800x800 matrix is used to reconstruct the original data component by component, it can be determined how much of the data variance is explained by each. Since the components gradually account for less and less of the variance, later components are of little relevance and can often be dropped with minimal loss of information. I will use this property of PCA to reduce the dimensionality of the data, as long as the resulting model still accounts for enough of the variance of the original data.

With this reduced brain response data, an attempt is made to find similarity based on semantic content of the image. Reshaping the principal components to a single column allows comparison by vector correlation. If one person's responses to animal images are more similar to each other than to responses to non-animal imagery, that would identify and localize an interesting relation between brain activity and semantics. Specifically, for the proposed game this might mean that a general semantic classification based on PCA of a brain response is possible.

The images are ordered by their semantic category, with images 1-400 being animal pictures and 401-800 non-animal. So if a matrix is constructed (see Figure 2) with the similarity in either an individual subject's or the average response between every combination of two of the images, an effective measure of response similarity is expected to show a clear difference per quadrant, with generally higher values in the animal - animal and non-animal – non-animal quadrants along the diagonal and lower correspondence in the quadrant with combinations between animal and non-animal imagery.

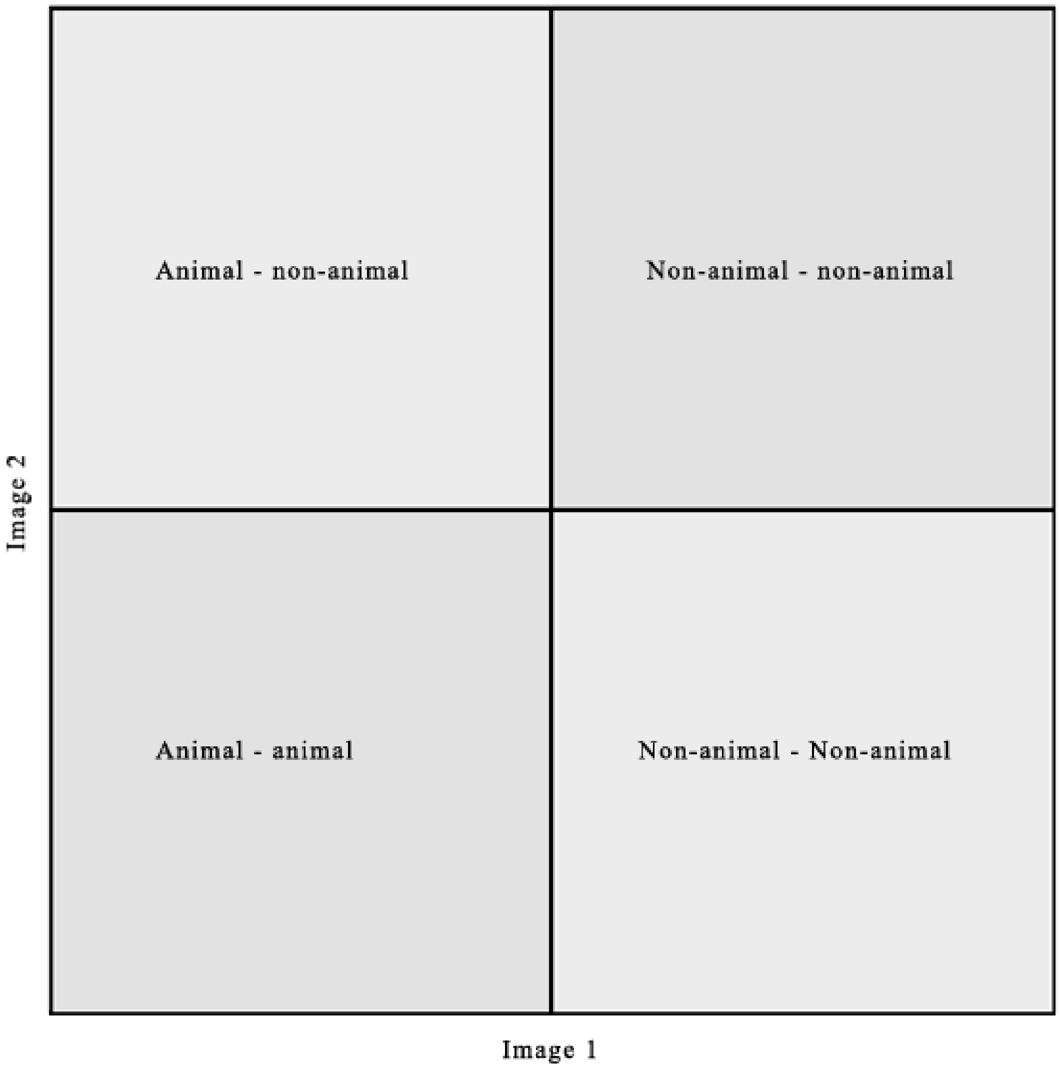

Figure 3: Values are expected to be bound to the quadrant. Note that correlation and least squares are symmetric functions, which means that the graph is symmetric in the x=y diagonal. Also, values on the diagonal are necessarily 1 (for correlation) or 0 (for difference-based methods). This will have to be compensated for when evaluating quadrants.

Another test that can be performed to check the effectiveness of a similarity calculation is to compare responses on the same image by different subjects. After all, that is closer to the objective in a two-player tagging game. This will yield a 33x33 similarity matrix. A good measure of similarity of brain activity should yield a matrix with generally high values of correspondence.

A final criterion of the method of response comparison is to see if an individual response can be correctly recognized by comparing it with the averaged response of all subjects. If implemented for every response, this will result in a 33x800x800 matrix. If averaged responses are columns and individual responses are rows, a correct classification of an individual response is made if the highest value in the row is in the diagonal.

4.2 Game description

If a sufficiently reliable way to match brain responses with image content is found, a prospective game could be constructed as follows. Unlike von Ahn’s game it would not be a publicly available internet game but conducted in a lab environment, because of the use of EEG readings. Players wearing EEG skullcaps would be randomly matched and play the game against each other. Each game starts with selecting “secret” images for each player and a random subset of the images in the database. The purpose for the players is to find out which of the set’s images is their opponent’s secret image, before the opponent guesses theirs.

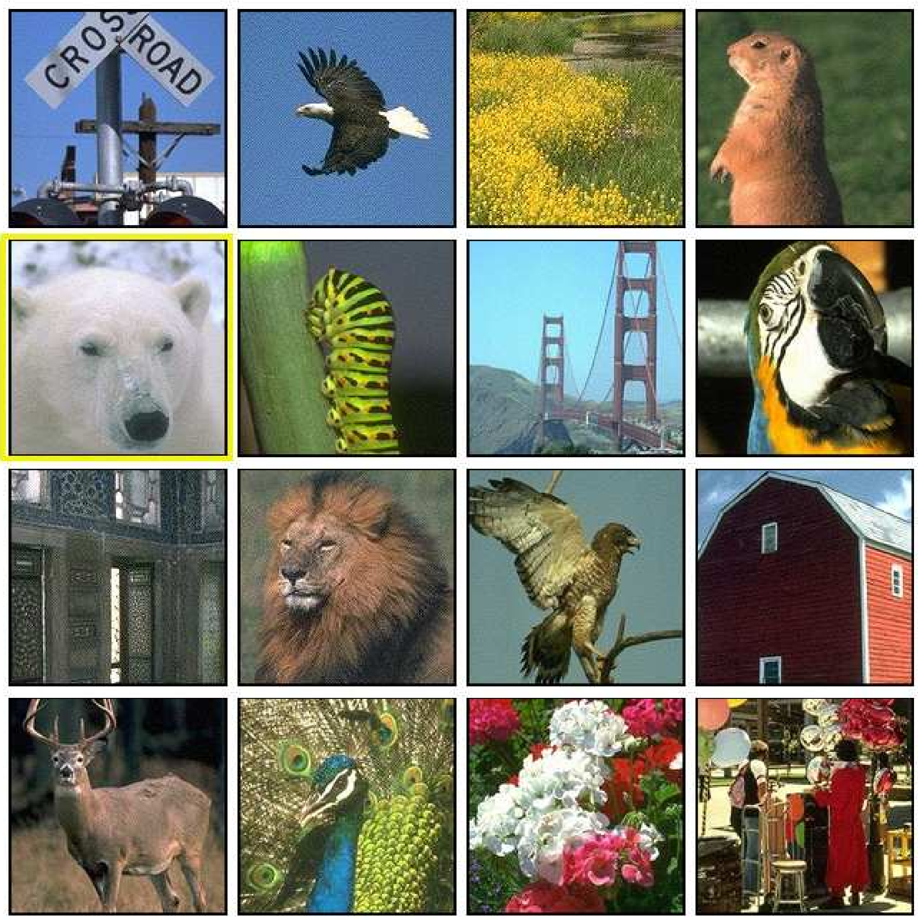

Figure 4: Players are presented with thumbnails of a number of candidates. Here the yellow outline of the polar bear image indicates that it is the target image of the current player.

The subset of candidates is shown by thumbnails. Players can select an image, and in turn choose to test or dismiss the image. If they choose to ‘test’, both players will be exposed to a full-size version of an image (instead of a thumbnail) under conditions similar to those in the earlier experiment by Ghebreab et al. The testing player will look at his chosen picture, while the other player is exposed to his own secret image. The similarity of the response of player A and player B is calculated and used as an interactive element to ‘score’ or grade performance. This can be shown by either a bar or a number;

a high value or large bar indicates that their response was highly similar. This gives the players a clue regarding how close they are to guessing the secret image.

Guessing the correct image rewards the player with a number of points depending on speed and similarity of the candidates (difficulty). Additionally, there are many possibilities to enhance this setup with entry of semantic tags. For example, players could be rewarded with test opportunities if they agree on enough tags for the subset of images. Or selection of images could be based on tags, imagine players enter “test lion” accompanied by a click if they want to test the match between their opponent’s picture and the lion picture.

Finally, if the similarity measure turns out to be not accurate enough or not always reliable, guided yes/no questions between the players could make it a bit easier, and it would allow additional tagging to be done. This can be imagined as one player posing a question like ‘Is it a bird?’ and if the other players replies negatively, the former can dismiss all bird pictures, effectively flagging each picture that does contain the semantic concept ‘bird’.

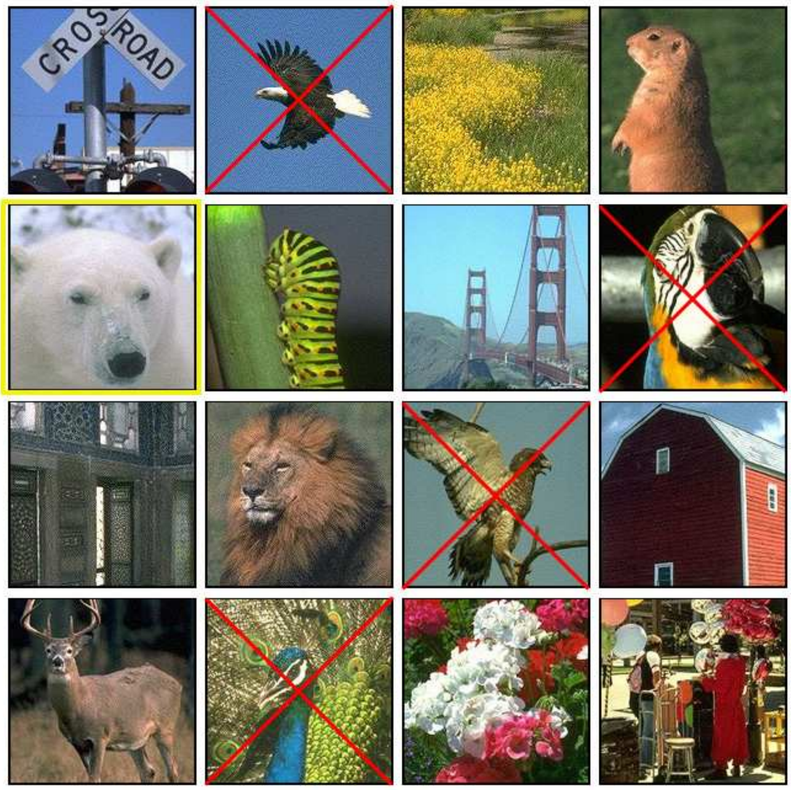

Figure 5: A player can dismiss images from the interface, thereby implicitly tagging images by their content. This way the game can collect both semantic tags and EEG data.

5. Results

First, principal component analysis was applied to every response. An individual sensor value $v_{S,I,s,t}$ of a response R is defined by test subject S and an image I, with sensor s and time index t. So, an entire response of one subject to every image is defined as the s by t by I matrix R_S. By reshaping the response to a single image $R_{S,I}$ to a column vector, R_S can be converted to a two-dimensional 13200 by 800 matrix. PCA is a transformation T applied separately to every R_S, so that $T(R_S)$ = P_S. The resulting 33 principal component matrices P are sized 800 by 800. But as shown in figure 5, our situation allows to reduce the number of these components from 800 to about 25 while retaining most information. This is a $1 - \frac{25}{800}$, or 93% data reduction compared to the individual sensor readings *v*. The relatively low number of required components means that there is a significant part of the original variables that did not particularly contribute to the distribution of variance by either being highly correlated or having a very low variance. An PCA per image was also performed, where $T(R_I) = P_I$. This yields 800 33x33 coefficient matrices.

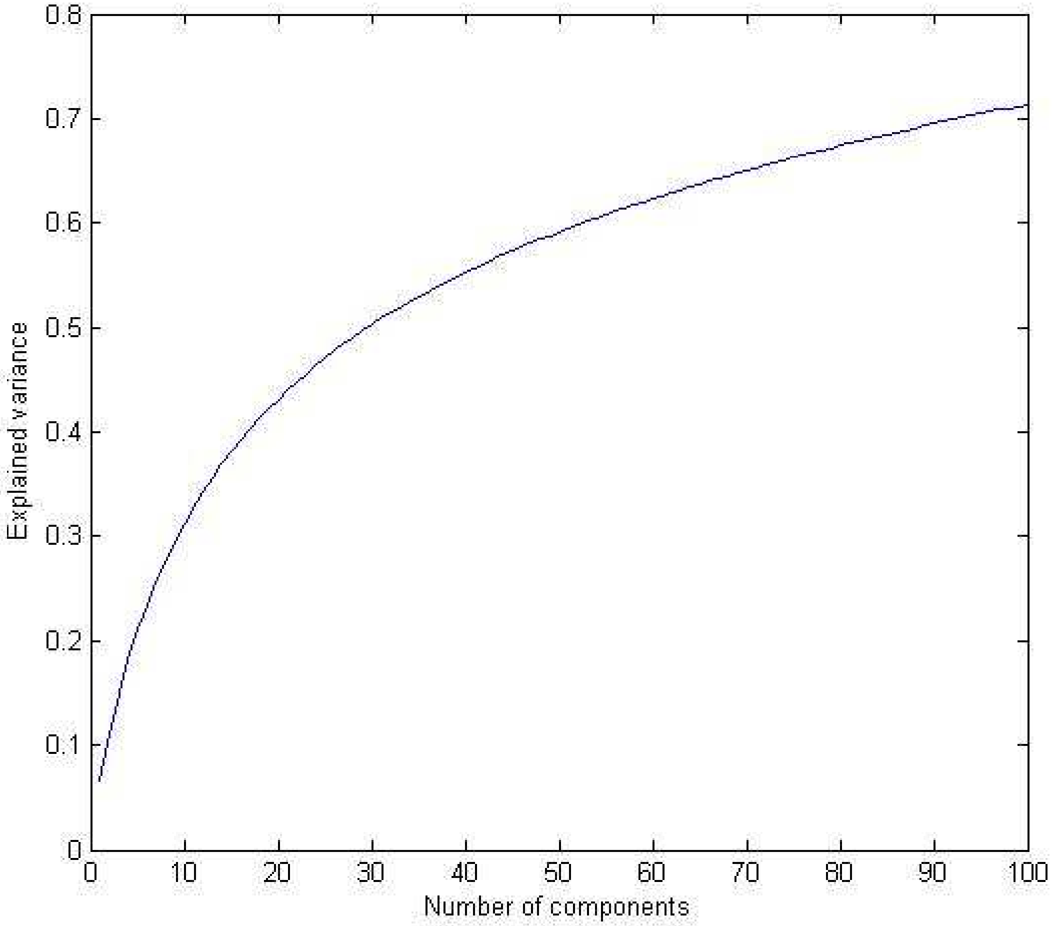

Figure 6: Graph of the relation between number of principal components and total variance accounted for. The first 30 components are sufficient to explain half of the variation in the data. Obviously the other 770 components are of relatively little importance. Arbitrary as it may be, I place the start of the factorial scree at 25. [16]

I started with examining the correlation between different responses to the same image. Euclidean distance between two PCA scores is used as a measure of distance. The values on the diagonals of the resulting 33x33 matrices are thus necessarily zero. I did not find very striking results here, although .

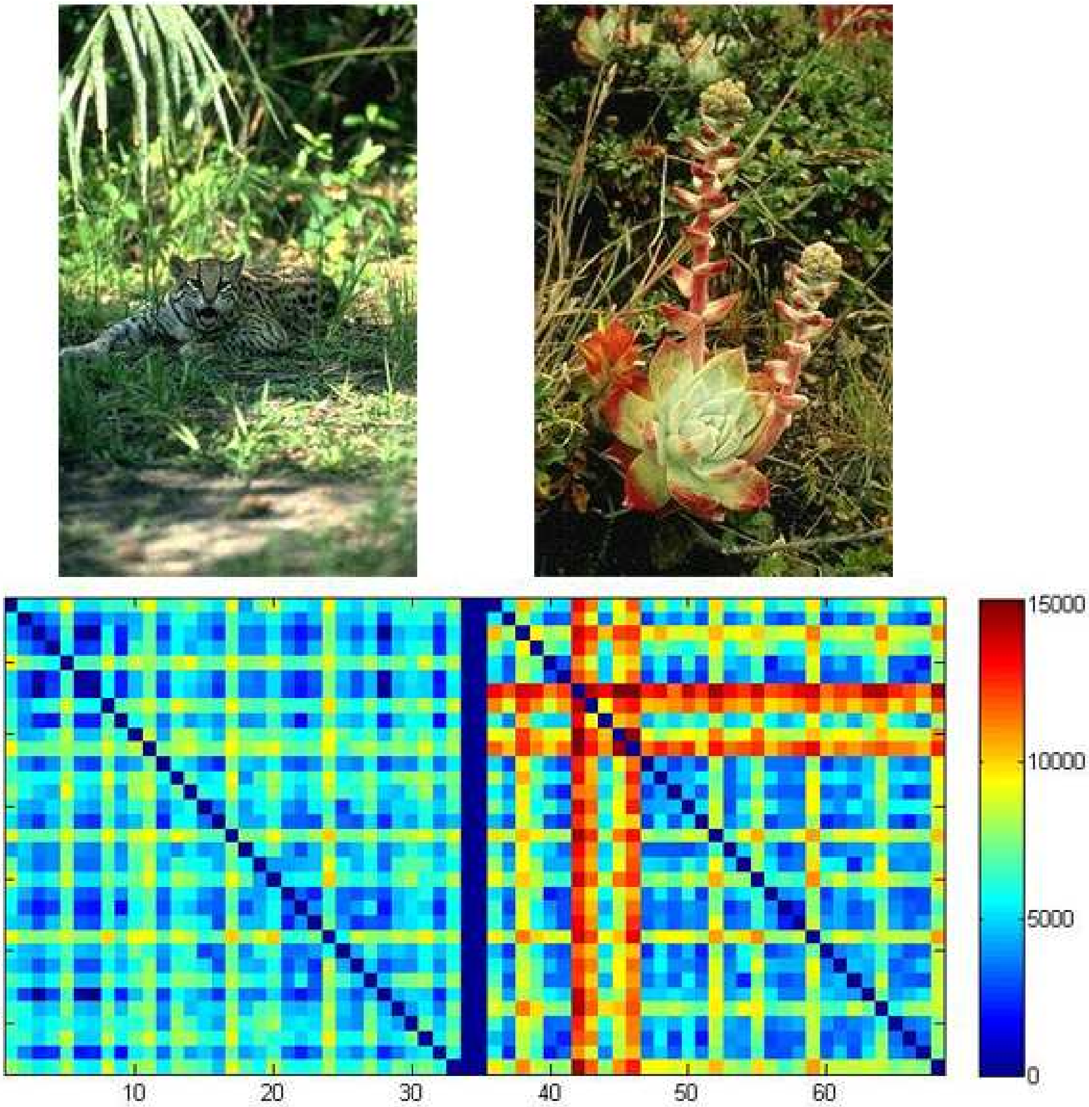

Figure 7: The feline on the left is the image that caused the most similar responses across subjects, while brain activity in on seeing the plant on the right was much more varied. However, the degree of dissimilarity for the lower seems to be caused by only a few subjects with very different responses, not by larger distance across the board.

But high similarity is not essential per se. If generally low values are adequate to correctly identify the image or even the general semantic category associated with a response, this would be sufficient to base the prospective game on. Therefore, let's see if our PCA-based method can correctly classify responses with some degree of accuracy.

For every test subject two new (dis-)similarity matrices are constructed. The first simply shows the Euclidean distance between his responses to different images, while for the second one, that distance is calculated between a subject's individual responses and the cross-subject average responses to the images. For single subjects, average distances for same-category images are slightly lower than distances between images of different categories. The difference is however rather small, with same-category images being less than 1% closer to each other. When divided by the semantic categories, as shown in Figure 2, the average and extreme resulting correlations are in tables 1 and 2.

Average Euclidean distance	Image A is Animal	Image A in Non-animal
Image B is Animal	5146	5203
Image B is Non-Animal	5203	5191

Table 1: On average, differences in distance based on semantic category are quite small

As a classifier of semantic category, average distance of response PCA with other responses of the same person achieves decent results, with 73.1% of the responses to animal images recognized (see Table 3). This is based on the first 25 components. Increasing that number can improve total accuracy to about 69% and animal recognition to 78%, but at the cost of a significant increase in computation.

	Image in animal	**Image in non-animal**	
Classified as animal	9657	5715	
Classified as non-animal	3542	7485	
Accuracy:	73.1%	56.7%	**64.9%**

Table 2: Using average correlation with a semantic category as classifier.

For this situation with 3 degrees of freedom, Pearson's chi-square yields an X^2 value of 3070. The p-value for the null hypothesis that the results of the classifier are not just uniformly distributed is less than 0.0001. Therefore, the null hypothesis is accepted, and the results can definitely be consider significant.

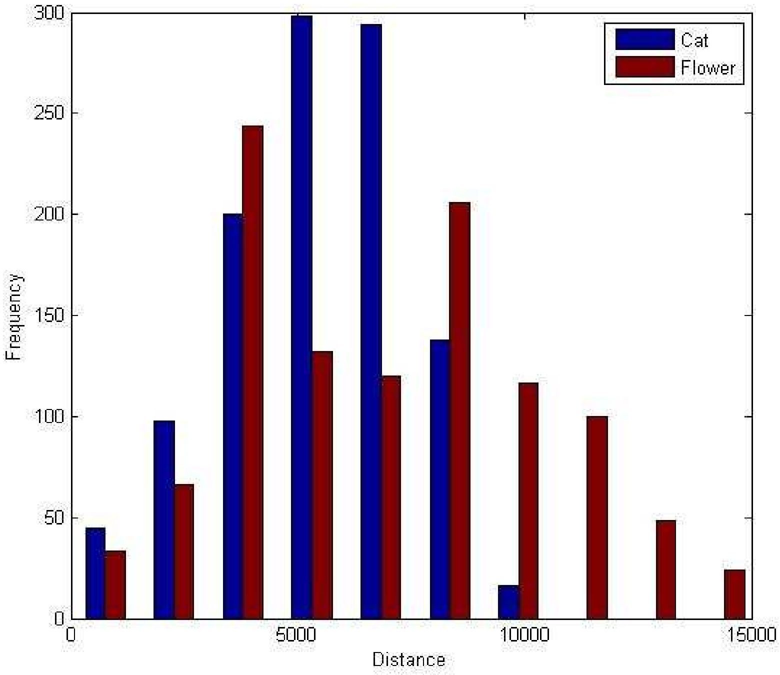

Figure 8: Distribution of Euclidean distances between pca scores for the cat and flower image from figure 6.

Figure 7 shows how response distances of the extremes of dissimilar and similar responses are distributed.

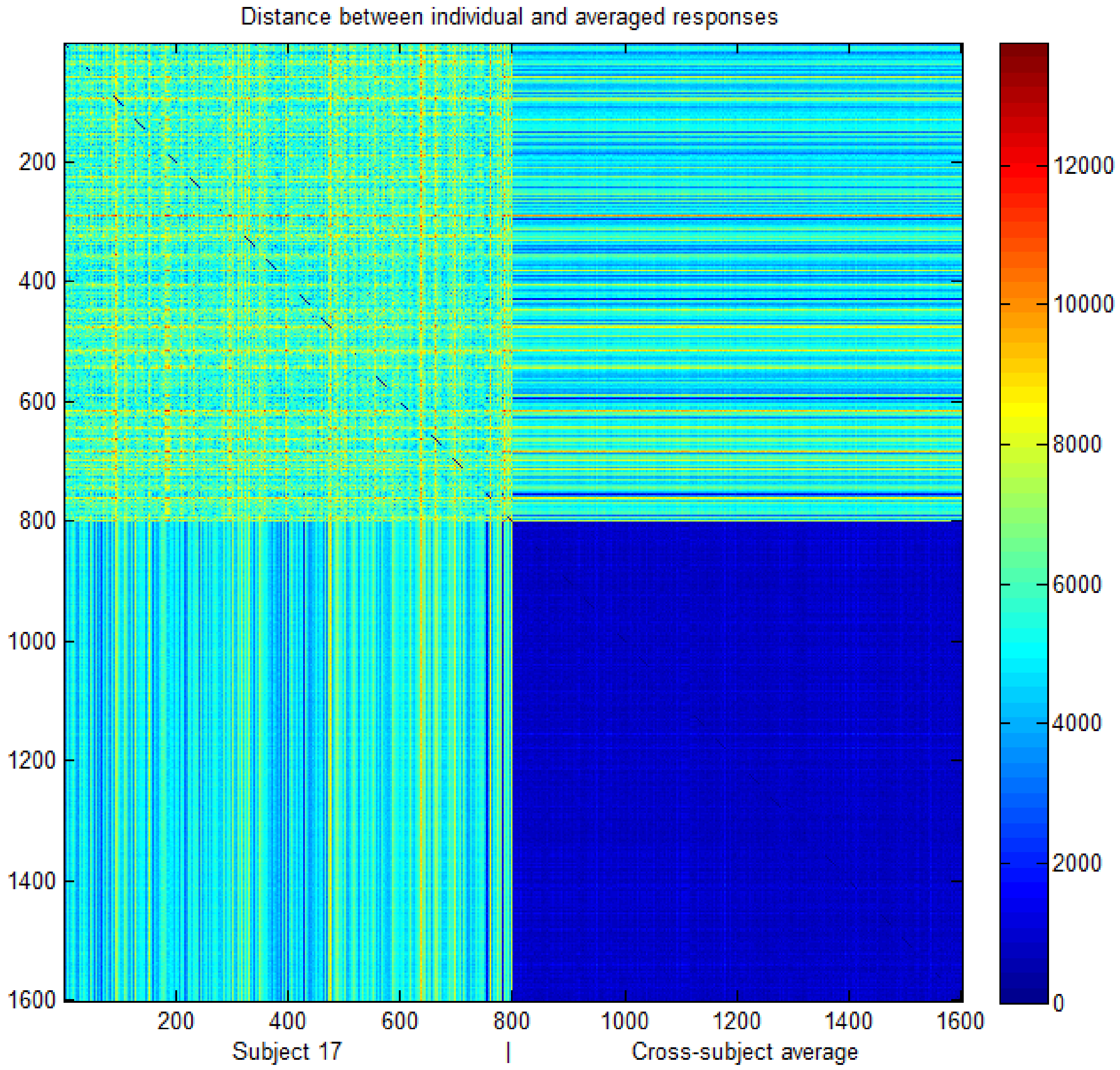

Figure 9. Distance of the PCA scores of both individual and averaged responses

Memory limitations restrict the computation of principle components across all images and all subjects at once. This makes it hard to simulate the classification of an unseen response to a specific image.

However, Figure 9 shows an attempt to do so. The cross-subjects averaged responses to every image were added as 800 extra rows to the matrix of one individual subject. Accordingly, the principal components of this large 1600x13952 matrix were computed. The figure shows a color-scale visualization of the distances between the scores of each of the two responses – individual and averaged - to each image. The top left quadrant shows individual to individual distances, the bottom right shows the relatively low distances between averages, while the remaining two quadrants show individual to average distance. It seems to be remarkably ineffective as a classifier: not a single one of the lowest distances per row in the quadrant is located where it matches correctly with the average of the same image.

6. Discussion and Conclusion

It has been established that using principal components of EEG responses can to some extent bring out semantic differences and similarities of the original images. I have specifically shown that Euclidean distance between principal components of a response is a useful classifier for semantic category. However, accuracy (around 70%) might not be high enough to be a sufficient performance method on its own for the game I have described. Its performance could possibly be improved so it will have a higher precision, or be complemented by additional measures that together produce the desired level of performance. On the other hand, a performance indicator that is just that, an indication and not a 100% related measure might exactly be what makes the game work, as a certain amount of chance rarely hurts a game,

It should also be noted that the ability to collect semantic tags and brain activity data does not necessarily require the similarity measure to be perfect, just workable, which it should be. The question of how brain activity data can be used to judge which exact image was seen, unfortunately remains unsolved. The way I have tried to measure similarity to detect specific images, individual static features of an individual’s personal brain activity appeared to be of much larger importance than likeness to the average.

The histogram of required principal components (Figure 5) shows that around 2% of the responses are almost completely explained by a single component. This seems to indicate that some values are missing or wrong. It is therefore possible that corrupt data has somewhat influenced the degree, but it seems unlikely that the difference it has made is relevant to the conclusion. Also, inspection of numerous cross-subject similarity graphs has given me the impression that it are always the same subjects who have different, outlying responses. See for example figure 6, where subjects 24 and 17 are separate from most others in both images. This is likely to influence classifying performance, but the reasons behind this are unclear and might warrant further investigation.

The relevance of this paper is thus based on the identification of a couple of approaches that do or do not work, as well as the description of the setup and requirements of a possible EEG-based semantic tagging game. It has been shown that EEG does indeed carry an amount of information on the semantic image content, and this conclusion could contribute to the development of an EEG-based game similar to the one I have described.

In the next paragraph I evaluate how the methods I have used could be improved to produce the desired results and I will discuss a number of other approaches.

7. Future Work

To improve the measures of similarity, a more thorough investigation of current results could already make a difference. Furthermore different parameters for averaging, measuring and other numbers of components could lead to slightly better results.

An option to improve explanatory power of the method is to look into a non-linear way of dimensionality reduction. The dimensionality reduction by PCA I used might not be sufficient, unable to capture certain regularities in brain activity data. Since brainwaves are inherently complex, periodical and nonlinear structures, it might be better to use methods specifically designed to deal with such data. With more advanced methods principal components could be calculated over all images and test subjects at once, finding other structures and similarities. For example, the continuous Weibull distribution [1] has been shown to be very useful in describing image content. Neural networks might also have potential to capture highly complex relations. An obvious advantage would be that a neural network is naturally well equipped to model a biological neural network.

For long-term future developments in concerning the semantic interpretation of brain activity, certain ethical questions must also be addressed. If it becomes possible to judge from brain activity what a person is seeing, or possibly even what he is dreaming or thinking, it is important to realize that what can be done to voluntary test subjects could also be used against someone’s will. In this case, this could theoretically have serious implications for privacy.

Of course there are probably as much or more excellent applications to think of. For example, a reliable method to interpret the semantic image content in brain activity could provide a useful tool for diagnosing psychological (schizophrenia) or visual disorders.

References

[1] Ghebreab, S. & Scholte, H.S. A Biologically Plausible Model for Rapid Natural Scene Identification (2009).

[2] Flickner et al (1995). Query by Image Content: The QBIC System. *Reading in Multimedia Computing and Networking*, p 255-263.

[3] Smith, J.R. & Chang S.F. (1997). Visually searching the web for content. *IEEE Multimedia Magazine*, Volume 4, Issue 3, 1997, p. 12-20.

[4] Daugman, J.G. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields (1989) Biomedical Engineering Vol. 36, issue 1, Jan. 1989, p. 107-114.

[5] Ruderman, D.L. & Bialek W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters* , Volume 73, issue 6, August 1994, p. 814-817.

[6] Van der Schaaf, A. & van Hateren, J.H. (1995). Modelling the power spectra of natural images: Statistics and information . *Vision Research*, Volume 36, Issue 17, September 1996, p. 2759-2770

[7] Hare, L., Lewis, P.H., Enser, P.G.B. & Sandom, C.J. (2006). Mind the Gap: Another look at the problem of the semantic gap in image retrieval. *Multimedia Content Analysis, Management and Retrieval 2006*, 17-19 January, p. 607309-1.

[8] Lew, M.S. Next Generation. Web Searches for Visual Content. *Computer.* Volume 33, Issue 11, November 2000, p. 46-53.

[9] Anderson, D.P., Cobb, J. et al. (2002) SETI@home: an experiment in public-resource computing. *Communications of the ACM,* Nov 2002, Vol 45 Issue 11, p 56-61.

[10] von Ahn, L. & Dabbisch, L (2004). Labeling Images with a Computer Game. *ACM Conference* on Human Factors in Computing Systems. CHI 2004, p. 319-326.

[11] von Ahn, L. & Dabbish, L. (2008). General Techniques for Designing Games with a Purpose. *Communications of the ACM*, August 2008. p. 58-67.

[12] von Ahn, L.(2006) Games with a purpose. *Computer*, Vol 39, Issue 6, June 2006, p. 92-94.

[13] Delorme, Rousselet, Macé & Fabre-Thorpe (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*, Volume 19, Issue 2, April 2004, p. 103-113

[14] Fize, Boulanouar, Chatel, Ranjeva, Fabre-Thorpe & Thorpe (2000). Brain Areas Involved in Rapid Categorization of Natural Images: An Event-Related fMRI Study. *NeuroImage*, Volume 11, Issue 6, June 2000, p. 634-643

[15] Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. (2008). Identifying natural images from human brain activity. *Nature* 452, p. 352-355

[16] Cattell, R. B. The scree test for the number of factors (1966). *Multivariate Behavioral Research*, 1(2), 245-276.