UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Rapid Serial Visual Presentation for Relevance Feedback in Image Retrieval with EEG Signals.

A Degree's Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Sergi Porta Caubet

In partial fulfilment

of the requirements for the degree in

Degree in Science and Telecommunication Technologies Engineering

Advisors:

Eva Mohedano & Noel O'Connor (Dublin City University)

Amaia Salvador & Xavier Giró-i-Nieto (Universitat Politècnica de Catalunya)

Barcelona, February 2015

Abstract

This thesis explores the potential of relevance feedback for image retrieval using EEG signals for human-computer interaction. This project aims at studying the optimal parameters of a rapid serial visual presentation (RSVP) of frames from a video database when the user is searching for an object instance. The simulations reported in this thesis assess the trade-off between using a small or a large amount of images in each RSVP round that captures the user feedback. While short RSVP rounds allow a quick learning of the user intention from the system, RSVP rounds must also be long enough to let users generate the P300 EEG signals which are triggered by relevant images. This work also addresses the problem of how to distribute potential relevant and non-relevant images in a RSVP round to maximize the probabilities of displaying each relevant frame separated at least 1 second from another relevant frame, as this configuration generates a cleaner P300 EEG signal. The presented simulations are based on a realistic set up for video retrieval with a subset of 1,000 frames from the TRECVID 2014 Instance Search task.

Resum

Aquesta tesi explora el potencial de les tècniques de *Relevance Feedback* utilitzant senyals EEG per interaccionar entre màquina i usuari. En aquest projecte s'estudia quins son els paràmetres òptims quan utilitzem *Rapid Serial Visual Presentation* amb *frames* procedents d'una base de dades de video. Les simulacions presentades en aquesta tesis mostren el *trade-off* que hi ha al utilizar un nombre petit o gran d'imatges en cada ronda del RSVP que captura l'interacció de l'usuari. Mentre que rondes petites del RSVP permeten un aprenentatge rapid de l'intencionalitat des del sistema, les rondes del RSVP han de ser al mateix temps suficientment llargues per permetre als usuaris generar la senyal P300 EEG que han estat marcats temporalment com a rellevants. En aquest treball també es fa referència al problema sobre com distribuir les potencials imatges rellevants i no rellevants en una ronda de RSVP per maximitzar les probabilitats de mostrar les rellevants com a mínim amb una separació d'un segon, ja que aquesta configuració genera senyals P300 més netes. Les simulacions presentades es basen en una configuració realista per recuperació d'imatges des de vídeo, treballant amb un subconjunt de 1000 imatges del TRECVID 2014 en el camp de Cerca d'Instàncies.

Resumen

Esta tesis explora el potencial de las técnicas de *Relevance Feedback* utilizando señales EEG para interaccionar entre máquina i usuario. En este proyecto se estudia cuáles son los parámetros óptimos cuando utilizamos *Rapid Serial Visual Presentation* con *frames* procedentes de una base de datos de video. Las simulaciones presentadas en esta tesis muestran el *trade-off* que hay al utilizar un número pequeño o grande de imágenes en cada ronda del RSVP que captura la interacción del usuario. Mientras que con rondas más pequeñas del RSVP permiten un aprendizaje rápido de la intencionalidad del sistema, las rondas del RSVP tienen que ser al mismo tiempo suficientemente largas para permitir a los usuarios generar las señales P300 EEG que han sido marcadas utilizando un *trigger* temporal como relevantes. En este trabajo también se hace referencia al problema sobre cómo distribuir las potenciales imágenes relevantes i no relevantes en una ronda RSVP para maximizar las probabilidades de mostrar las relevantes como mínimo con una separación de un segundo, ya que esta configuración genera señales P300 más limpias. Las simulaciones presentadas se basan en una configuración realista en recuperación de imágenes desde video, trabajando con un subconjunto de 1000 imágenes del TRECVID 2014 en el campo de Búsqueda de Instancias.

Acknowledgements

First of all I would like to thank my thesis advisors at Universitat Politècnica de Catalunya, Xavier Giró-i-Nieto and Amaia Salvador, for their support and advice during my project.

During the thesis, I worked at the Insight centre for Data Analytics at Dublin City University. I would like to thank them for that chance and to let me do my research using all their tools and resources. Especially to Noel O'Connor who was my advisor at Dublin City University.

The main support during this thesis was the Phd student at DCU Eva Mohedano who has been implied in all the parts of the project as it was a continuation of previous studies she did in her Bachelor's Thesis and also during her Phd.

Revision history and approval record

Revision	Date	Purpose
0	1/01/2015	Version 0
1	26/01/2015	Version 1
2	05/02/2015	Reviewed by Xavier Giró.
3	05/02/2015	Reviewed by Sergi Porta
4	06/02/2015	Reviewed by Xavier Giró
5	06/02/2015	Reviewd by Sergi Porta

DOCUMENT DISTRIBUTION LIST

Name	e-mail
[Student name] – Sergi Porta Caubet	sergiportacaubet@gmail.com
[Project Supervisor 1] - Xavier Giró I Nieto	xavier.giro@upc.edu
[Project Supervisor 2] – Noel O'Connor	Noel.OConnor@dcu.ie
[Project Supervisor 3] – Eva Mohedano	evamohe@gmail.com
[Project Supervisor 4] – Amaia Salvador	amaia91@gmail.com

Written by:		Reviewed and approved by:	
Date	06/02/2015	Date	06/02/2015
Name	Sergi Porta Caubet	Name	Xavier Giró i Nieto
Position	Project Author	Position	Project Supervisor

Table of contents

List of Figures

List of Tables:

1. Introduction

1.1. Statement of purpose

Interest in the processing of video has increased enormously during the last few years. Users in many fields are exploring the possibilities of manipulating visual data in many different ways. However, due to the huge amount of available content, has appeared the need for searching and filtering content of interest in an efficient way. The problems and solutions focused on the retrieval of images taking into account some features as color, texture and shape have generated a field in the technologies world called *Content-Based Image Retrieval (CBIR).*

A branch of this research has focused on the retrieval of video based on visual queries. Given an object of interest, there are some algorithms that try to retrieve images containing instances of that object in a video collection. However, there is a trade-off between the retrieval speed and its accuracy. One way to improve both is through human interaction, asking the user to provide feedback about the relevance of the some results selected by the algorithm.

Even the best ‘visual based’ search engine may have problems in understanding what is the user intention, as this is stored in the brain. The only way to know the exact user intention is asking him and, this is, involving the human user in the retrieval loop. It has been proved that involving users increases the quality of the ranked list of results. The problem lies on the need of annotation inputs in each iteration to train the models, because it requires a human to look at those images and a computer to capture some kind of feedback interaction. From this point on, there are two main discussions: which model reaches better performance and how we train these models. This project will firstly focus on introducing smart selection of the visual content to be shown to the user, proposing those images which can help more in the search of visual instances on a large video collection, such as objects or people.

It has been proved that to process visual content, the human visual system is the most powerful tool. Our brain is able to process an image in less than a few *ms* with an enormous accuracy but the problem appears on its performance. On the other hand, computer vision systems are powerful and scalable in terms of computation, but still lack accuracy in image understanding. Moreover, the capability of our brain is nowadays still superior in that sense, and it is able to extract semantic content of an image, like it could be objects, people or other features. These cognition processes generate electrical signals in our brain and some of them can be measured and analyzed. So there is a research field that aims at interpreting these signals triggered by the human vision system and, in particular for the scope of this thesis, on electroencephalography (EEG) signals measured on the human scalp with a non-intrusive Brain Computer Interface (BCI).

So the there is a huge amount of possibilities since the brain takes part on it. The main point is how to combine the power of the human visual system with the image retrieval algorithms and even more important, how to improve them.

The project main goals are:

1- Reduce the user interaction in image retrieval by introducing an active search approach that will allow a small selection of the images to be annotated as relevant or non-relevant.

2- Explore the potential of Brain Computer Interfaces to solve an instant search problem from a large video collection.

1.2. Methods and Procedures

This project is a continuation of Eva Mohedano's scientific publication [1] who studied how to use brain signals to cut object from images. Her studies combined BCI and computer vision focusing primarily on object segmentation from an image. Also we used the Matlab library called EEGlab, that provided all the tools we needed in a visual interface to allow us to process the brain data from the 31 BCI channels and to understand the results.

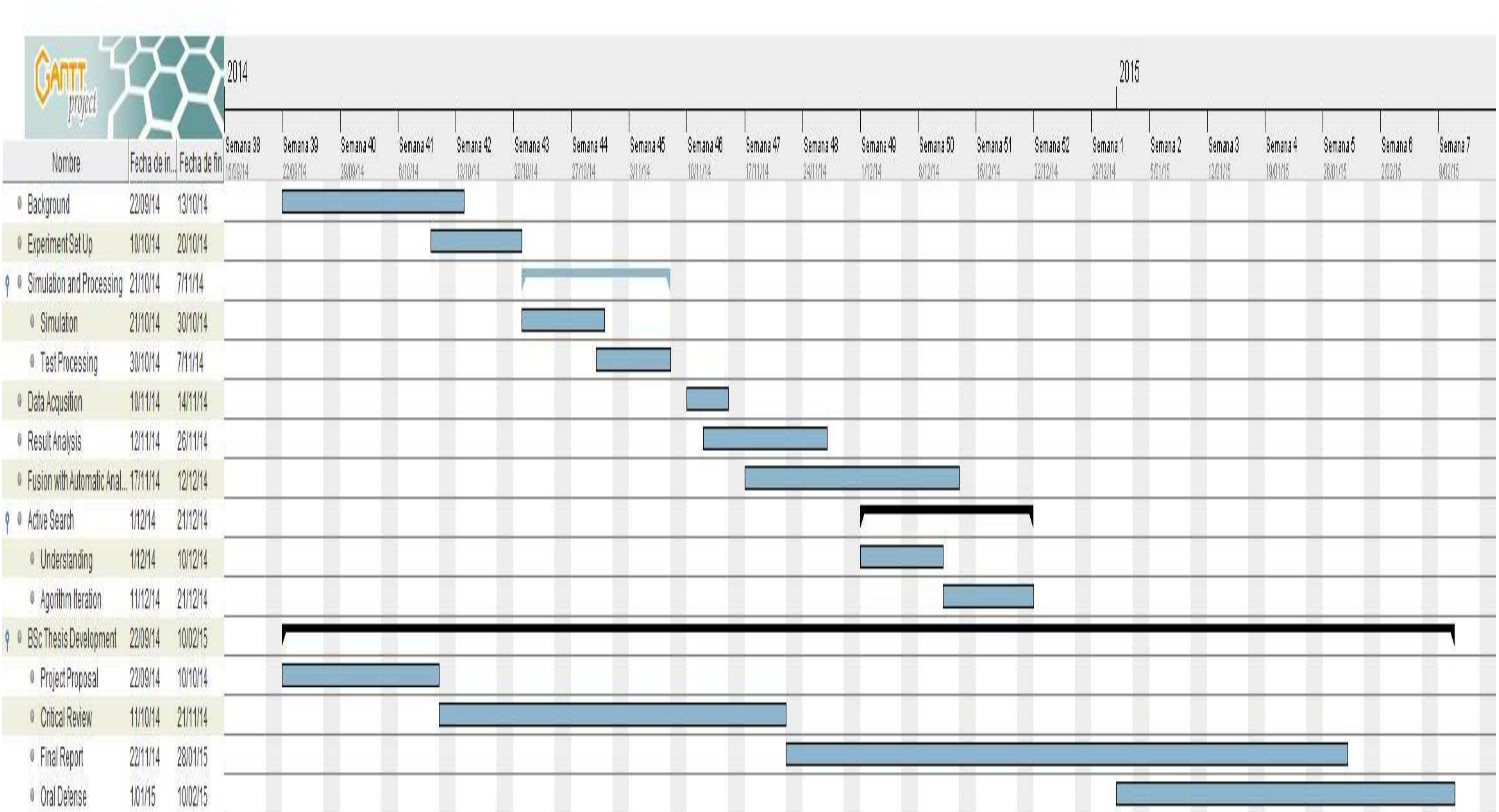
In the part of relevance feedback, we had to adapt to our data some scripts in Python that were already implemented as a part of Amaia's Salvador scientific publication [11]. From this we used a script that implemented the Query Expansion technique and also we used the Python library Sckit-Learn[1] to implement the Support Vector Machine and the K-means (Clusters) algorithm.

[1] http://scikit-learn.org/stable/

1.3. Workplan

In the annex there are the workpackages already showed in the Project Proposal and Critical Review documents with the description of any changes they had.

1.3.1. Gantt Diagram



1.4. Incidences

The first idea of this thesis was to annotate images using brain signals and then compare the results with the mouse interface. After collecting some data from some real users and checking that it was actually possible to annotate images with BCI, we found more interesting to point the project to the Relevance Feedback part and then try to relate it with the first idea. The comparative study of BCI vs Mouse interface was run in parallel by my co-advisors and, while I participated in the study, that was not the main task related to my bachelor thesis.

2. State of the art of the technology used or applied in this thesis:

This thesis is based on some previous works described in this chapter:

2.1. EEG Signals for Computer Vision

The first resource of information is from a paper[1] of the Dublin City University and Universitat Politècnica de Catalunya, carried mainly by my advidsor, Eva Mohedano. It talks about the usefulness of the brain computer interfaces in segmenting objects from images. It consists in splitting the image in smaller blocks that are displayed in a brain interface screen in order to get the EEG signals. Then, taking into account the intensity of these signals, it creates a probability map that is binarized and combined with vision computer segmentation algorithms it is able to cut the object from the image.

It worked splitting the image into little squares and showing them consecutively at a high rate. That technique is called Rapid Serial Visual Presentation (RSVP) and it consists in displaying some “target” images among most of distractors. In our case is exactly the same case but instead of displaying little parts of the image, we displayed whole images. The presentation rate of the images is high, between 5 and 10Hz, so the signature in the corresponding EEG signals is produced when the user observes the target images. This signature is known as P300 wave and it is a kind of Event-Related Potential (ERP) associated to the process of recognising a relevant visual stimulus. The wave’s primary characteristic is a positive peak in the EEG signal around 500ms after the visual stimulus is observed

It is important to focus on the data acquisition of this study because my project will carry on a very similar way to get the EEG signals. The difference lies on the kind of images displayed, because my target will be a whole image containing the object of interest and the distractors will be the images without the object. In that previous project, its targets depended on whether they were image blocks containing a part of the object or blocks without pixels of the object. Even though, the information extracted from the brain computer interface and the image display procedure will be very similar.

On the other hand, this project was also based on the image retrieval techniques applied by researchers of Columbia University[4]. It does a close study to our approach, trying to state how a brain computer interface can add accuracy to the computer vision algorithms in the annotation images field. It does a first trial of image retrieval using EEG signals and then with these signals it feeds an annotation algorithm to choose the images.

In this paper[4] it is shown how to understand a single trial of EEG decoding. It explains how we have to combine the different electrodes signals from the EEG interface in order to get a variable that describes whether an image is a target or a distractor. This paper also demonstrates how we can discriminate the noisy sample, by setting thresholds to determinate the tags. During all this process we will take into account the brain time response and the relevance of each impulse.

2.2. Humans in the Retrieval Loop

I based the final part of the project on the previous approaches of Relevance Feedback for image Retrieval[3] and especially on how to apply them to a Support Vector Machine with active learning[2].

The first paper[3] is important to understand the concept of Relevance Feedback, where the results retrieved in the further events take into account the information of the previous results in order to improve the accuracy. In the Standford paper [2] it explains how to implement a Support Vector Machine and how to add feedback on its process. So they study the different configurations that might work, studying the effect of different block sizes on each round and different annotation budgets and their effects on the accuracy of the final images list retrieved. So finally, these previous studies will guide us to implement a Support Vector Machine with Relevance Feedback in our images database that will let us know what could be the best configuration to combine Relevance Feedback with an EEG Annotation System.

Relevance feedback interactively determines a user's desired output or query concept by asking the user whether certain proposed images are relevant or not. For a relevance feedback algorithm to be effective, it must get the query concept accurately and quickly, but also only asking the user to label a small number of images. So if we export this concept to our context, we would like to know the minimum number of images that should be annotated to get a result with enough accuracy.

The main idea of the relevance algorithms is to start with a set of images with relevant information, and this way create a boundary for each query that quickly separates the target images from the rest of the dataset.

In the figures 1 and 2 there are shown some results from the Standford paper [2] where they state the efficient results they got when using Relevance Feedback. In those experiments, they labelled 20 images per query.

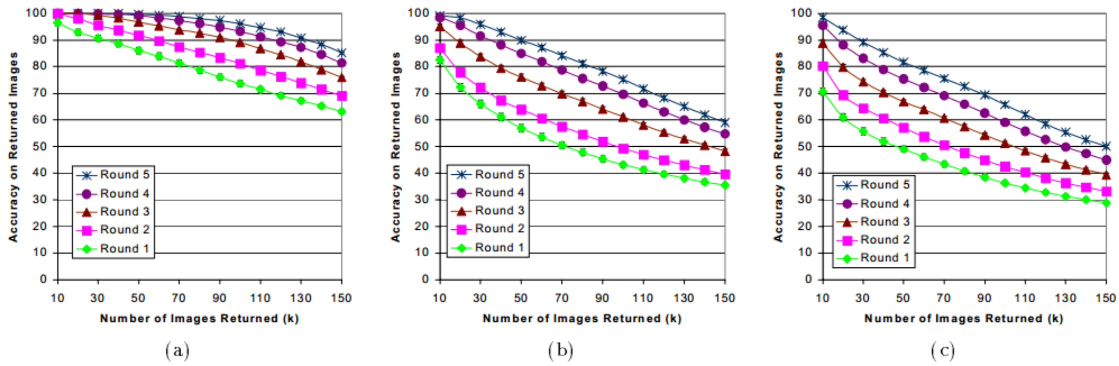


Figure 4: (a) Average top-k accuracy over the four-category dataset. (b) Average top-k accuracy over the ten-category dataset. (c) Average top-k accuracy over the fifteen-category dataset. Standard error bars are smaller than the curves' symbol size. Legend order reflects order of curves.

Figure 1. Different number of iterations versus retrieved images.

In the Figure 1 [2] we can see the accuracy of the three different categories. It decreases in function of the number of images returned, but if we focus on a k constant output images, then we can see that the accuracy improves on each round of the algorithm as we expected since we are increasing the number of labelled images in each round. So it is proving that the more iterations has the algorithm, the better the performance.

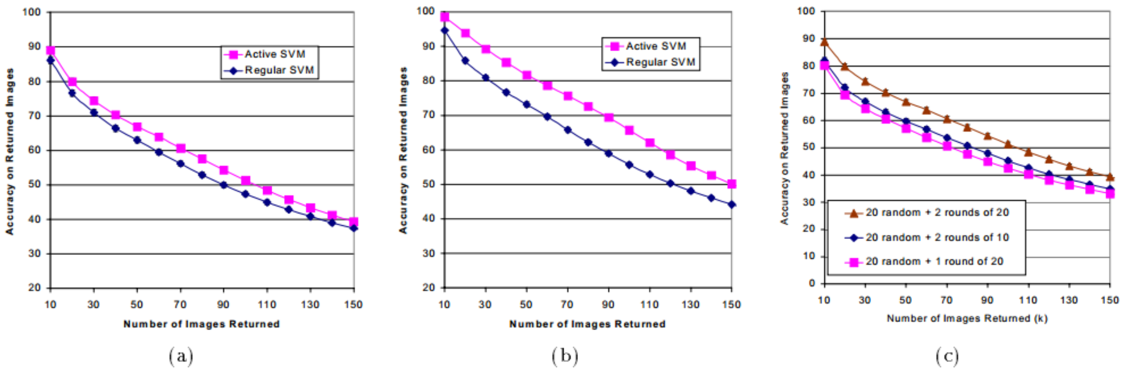


Figure 5: (a) Active and regular passive learning on the fifteen-category dataset after three rounds of querying. (b) Active and regular passive learning on the fifteen-category dataset after five rounds of querying. (c) Comparison between asking ten images per pool-query round and twenty images per pool-querying round on the fifteen-category dataset. Standard error bars are smaller than the curves' symbol size. Legend order reflects order of curves.

Figure 2. Regular SVM vs Active SVM.

In the Figure 2 [2] we can see that there is a significant increase of the performance from using the active method. The most relevant point here is in the (c) graph, where we can see that it is comparing two different strategies but labelling the same number of images. Its conclusion is that if we increase the number of rounds while decreasing the number of images per dataset, we achieve a little better performance.

3. Methodology / project development:

3.1. Relevance feedback algorithm

We have explored two different strategies to choose the images shown to the user so s/he can provide relevance feedback. The main goal is to annotate the minimum amount of unlabelled images and obtain the highest accuracy possible. This section explores two approaches to exploit the annotation in one round to select the images for the next round.: query expansion and a support vector machine classifier.

In both cases, the decisions are based on visual features automatically extracted from the whole dataset. These descriptors were extracted from the deep learning software Caffe[2] to generate vectors of 4,096 dimensions to represent the colours, textures and shapes contained in each image. This descriptors have been proved to provide a state of the art performance for image retrieval scenarios [12].

.

3.1.1. **Query Expansion**:

Here we will explain the main steps to implement the Query Expansion algorithm[11]. As we said before we have four query descriptors to fit the classifier in the first iteration as previous information. In the query expansion strategy, the data is structured in two different pools: one for the previous query descriptors and another one for the unlabelled frames descriptors.

To do that, by using the frame descriptors we compute each distance to the four query descriptors, and then we retrieve the ones with less average distance. So the steps are:

1- Start of the algorithm. We compare all the frame descriptors with the four query descriptors and compute the distances. Then we get a matrix of (1000x4), where each data is the distance between the descriptors.
2- For each frame row, we get the minimum value related with a single query descriptor. So from this point on, we just have a single distance related with each frame descriptor.

[2] http://dx.doi.org/10.1145/2647868.2654889

3- Then we sort the frames based on the distance in order to have the expected more relevant frames in the top of our column.
4- Finally we annotate the first N frames in the top of the sorted column by using the ground-truth file that gives us the label for each frame. In the next iteration the positive frame annotations are included in the query descriptors pool, while the non-relevant annotations are not used in the feedback process.

3.1.2. **Active Support Vector Machine**

To run this strategy we user a Support Vector Machine as a classifier[2]. The main idea is the same as in the query expansion strategy. We want to annotate a pool of unlabelled data using a classifier in order to know which images are more probably relevant, and also we want to get a good accuracy while tagging the minimum number of frames. The main difference lies on the classifier, as it takes information from the previous rounds using the relevant and the non-relevant frames.

The implementation steps in this case are:

1- Start of the algorithm. In the first iteration, as we only have positive annotations in the queries pool, we compute the distances the same way as in the query expansion strategy.
2- Then from the second round on, we use the SVM Classifier. The classifier is fit with the relevant and the non-relevant annotations from the previous rounds and it creates a hyper-plane with all the descriptors vectors.
3- In each round it compares all the frames in the unlabelled pool with the updated hyper-plane, and retrieves the frames which have the descriptors closest to the hyper-plane. Then these frames are annotated and fit the classifier again.

3.2. <u>Composition of relevant and non-relevant images per round</u>

In this point, we already have two different systems to retrieve images basing on some previous information. Once we decide which of them work better, we can make them improve in other ways. As we said the main background of that thesis lies on the idea of annotating images using EEG signals. The thing is that in order to use the human visual system to annotate images, we have to take other things into account. Like the frequency of image displaying and the separation of the targets to make sure that the brain can notice them.

After getting the sorted list from the classifier, it makes sense that if first we annotate the N images with highest score we get better results. The problem is that it would not be realistic in our EEG context as we can't display target images together in the EEG displaying. So we realized that we could introduce another feature to be studied in order to try to get almost the same results.

So from this point on, we will try to know how which images from the sorted classifier list we should annotate. As we will explain forward, the EEG interface works good when it has less than 25% of relevant images in each round. We studied some different compositions.

In this section we will be referring to the scores list. That list is the one retrieved from the classifier for each round. So on the top of the list there are the frames that the classifier thinks that are more similar to the query descriptors and at the bottom of the list the less relevant frames.

3.2.1. **Highest Scores**

This is the best distribution a priori. In that distribution we will be annotating images from the top of the scores classifier list. As we are getting always the images with more information and at the same time we are fitting the classifier with the more relevant images, it makes sense that we will improve results faster than with other distribution strategies.

Figure 3. In each iteration the frames with highest scores are picked to be annotated.

3.2.2. Random

In the random distribution we keep the original list without sorting it until the end of the experiment. From that list we annotate image sets that are completely random and after annotating each group, we compute the map sorting the relevant images found on the top of the stack and the non-relevant to the bottom.

3.2.3. 5-95% (EEG)

We called that one 5-95% because, as we will explain forward in the EEG experiments, there will be 5% of relevant and 95% of non-relevant images in each RSVP round. This configuration has been included because it corresponds to the experimental set up run in parallel, using blocks of 200 frames composed 10 relevant and 190 non-relevant images.

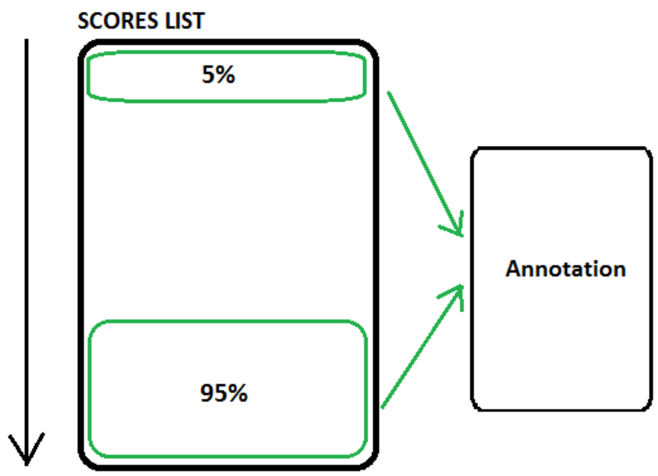


Figure 4. We choose 5% of the frames from the top of the scored list and 95% from the bottom of the list.

3.2.4. Clustering

That is the last distribution purpose and at the same time the one more sophisticated. It aims at including the maximum diversity in each round. To do that we use the k-means algorithm [3]already implemented in the Scikit-Learn Python library. That algorithm clusters data by trying to separate samples in groups of equal variance and also requires the number of clusters to be specified and then it tags each sample with a cluster label. In our case the number of clusters will be the size of the block in the iterations.

In the first iteration, the k-means algorithm labels all the frames in our database to different clusters and also retrieves a "centroid" for each cluster. The value of K must be specified and it corresponds to the amount of images to be included in each feedback round. Then using a method already computed in the clusters library, we can get the frame with minimum distance to that "centroid" in each cluster. Then we annotate these frames with more information and fit the classifier.

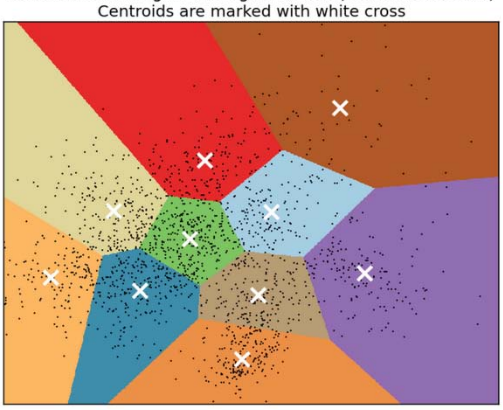


Figure 5. Clusters illustration. The white crosses indicate there the centroids are placed. The closest frame to each centroid is retrieved to be annotated.

3.3. <u>Implementation Details</u>

This relevance feedback experiments are run with Python language and using the Canopy software development kit. Python is a simple and minimalistic language that allows us to understand the code easily by using comments and compact structures. Also it is important to notice that Python is an open source, so this way we are allowed to distribute our code or use parts of other experiments. Another important feature is that it

[3] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

is a high-level language, so we do not need to bother about memory managing or other low-level problems.

In the annex will be attached all the code used to run the experiments as well as a high level description of it.

Our experiments have been based in some others scripts of our research group who[11] tried to retrieve a set of N images from a database using image descriptors. It is important before get into the technical parts of the algorithms, to explain where the images come from, as we will be dealing with them.

The image descriptors are created to have the images in a vector form, and this way it is easier to compare them. To create the descriptors it is taken into account features as colour, texture or brightness and then vectors of 4096 positions are created with all the different features. It is important to deal with the vector form of the images as if we wanted to compare them with their pixel level information, it would be almost impossible for the computation. The descriptors were extracted from the deep learning software Caffe[12].

4. Results

There are some main things that we want to figure out by running the different experiments:

1- Which feedback strategy works better to fit the classifier?
2- What distribution strategy should we use to adjust with the EEG experiments?
3- What is the optimal round size per iteration?
4- How should we sort the data to be realistic in the EEG display?

4.1. Experimental Set Up

4.1.1. Dataset

Our database is made of videos used in the TRECVID 2015 Instance Search task, which contains videos from the Eastenders. Actually, only 1000 frames from that collection were used in our experiments, containing 50 relevant frames among them.

In all of our experiments we have been working with four different target queries, and all the results are the average of the different queries. TRECVID defines each query with four images so the images selected in the first round were selected by using these four images as relevant.

4.1.2. Brain Computer Interface

A non-invasive 31 channels BCI with sample rate of 1KHz is used to capture the brain reaction of the users during the image presentation. The electrodes were located according to the 10-20 system distribution and the experiment was run in a Faraday Cage. This kind of room isolates the user and equipment to minimize the possible interferences from any other unrelated acoustic or visual events.

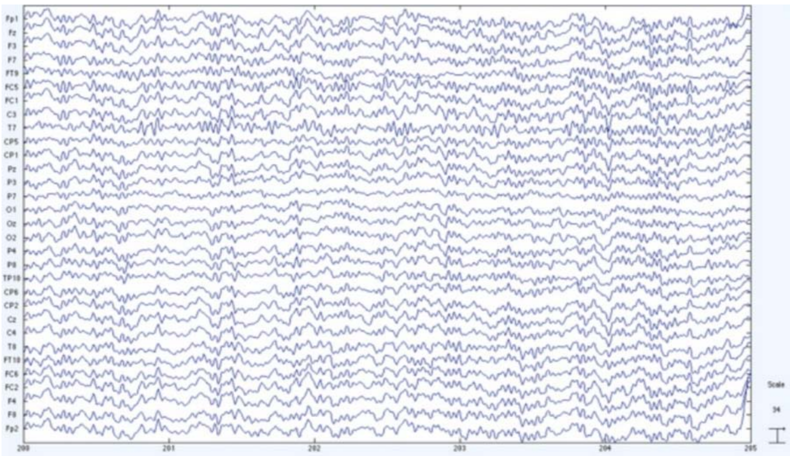


Figure 6. 31 BCI channels time response.

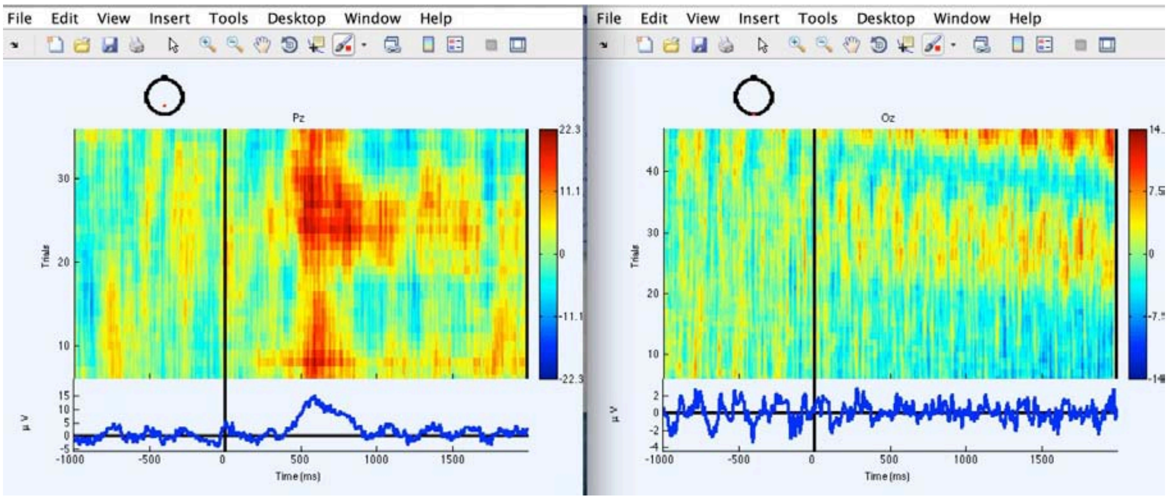


Figure 7. The figure at left represents the average response for all the query targets. The figure at right represents the average response for the distractors.

4.1.3. **User tasks**

Images presentation for each query in the experiments was carried out as follows. First, four different images containing the goal object were displayed. This allows the user to memorise the visual features of the possible images containing the object. The user had as much time as needed and the presentation started after pressing any key.

It is proved that participant attention decreased with time, so we asked them to press a key for each target image displayed. Also it helped us to have a first idea of how many targets the user got to see.

For each query we split the database into five blocks of 200 images. Each block contained 10 relevant frames against 190 distractors. The distribution of the target images among the rest was random. To know in the log files when the target images where displayed, we tagged them using the display software.

4.1.4. **Data post-processing**

The data was referenced to the Tp9 channel and subsampled from the original 1000Hz rate to 250Hz, as we did not need that resolution. Depending on the users, we had to use the Tp10 channel if the Tp9 was noisy. Then, a band-pass filter from 1Hz to 70Hz was applied. By visual inspection, we rejected manually the noisy segments. With the data filtered, we extracted the brain reaction related to the stimulus by selecting one and two seconds pre and post window presentation.

For the feature selection, we selected the time region within the epoch that best characterized the difference between targets and distractors. This region was contained between 200ms and 900ms after the visual presentation. The feature vectors were built by concatenating the 31 channels for this time region. The final feature vector was obtained by applying a second subsample to the vectors to reduce the sample rate to 20Hz.

4.1.5. **Evaluation Metric: Mean Average Precision (MAP)**

It makes sense to compare the two different strategies we need a precision measurement. In our context we will be always annotating the whole set of 1000 images for each query, where there are relevant and non-relevant images. So we need to find a measurement system that takes into account the amount of positive tags that we have at the beginning, rather than computing the accuracy of positive tags in the whole 1000 set, as we would never get the maximum accuracy. In the TRECVID benchmark, the metric used was MAP, so that is why we adopted it.

The MAP measurement takes into account how many relevant tags are at the beginning of the list. It means that if we have the first 50 positive images and then the rest are all negative, we get the maximum MAP value. It is always between 0 and 1, where 1 is the maximum.

As we said, we will be annotating always the whole database, so after knowing where the positive images are, it is logical that at the end of the annotation we will always get 1 as MAP value. The point here is to try to see how fast the curve goes to the maximum without annotating all the images and then decide which strategy is better.

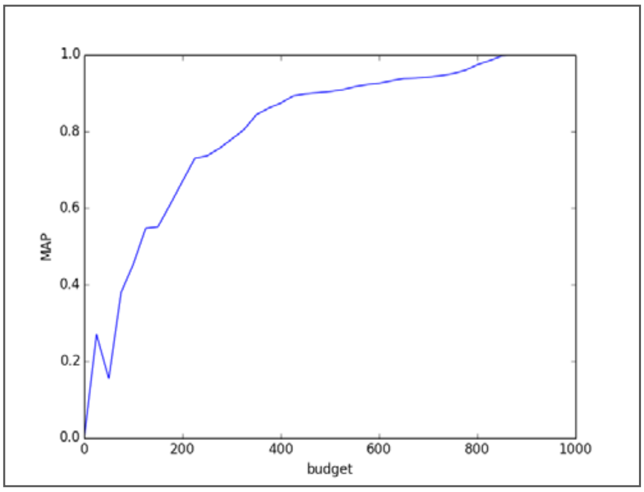


Figure 8. Mean Average Precision.

4.2. Relevance Feedback Strategy

First we wanted to know which relevance feedback strategy would get better results and which one we should apply to solve the next questions. So first we ran some experiments only using the Query Expansion strategy to study its behaviour.

The first thing we had to know is if we were improving the system when introducing more relevance feedback information. So we ran the experiment with different round sizes for the QEX strategy.

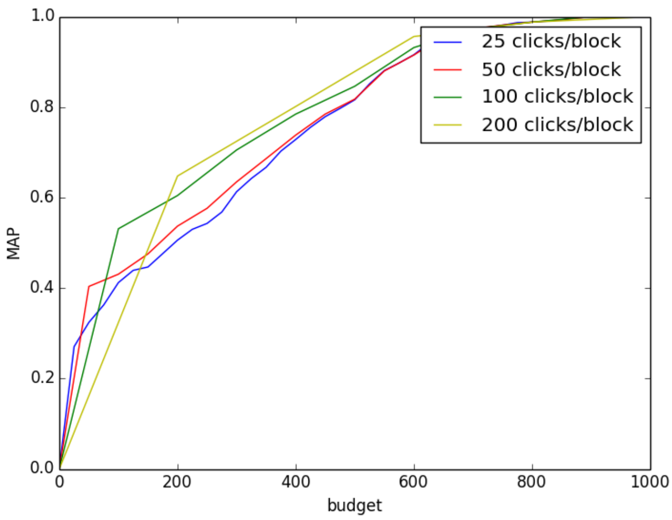


Figure 9. Different block sizes in the Query Expansion strategy.

What we would expect using relevance feedback is to increase the accuracy when increasing the number of iterations. In the Figure 9 we can see the query expansion results using different round size configurations annotating a database of 1000 images per query. So we ran the same experiment using four different round sizes from 200cliks/round to 25cliks/round.

What it does for each round is to add the annotated queries to the queries database to try to improve the searching of relevant images in the next rounds. If we consider that point, we would expect to get better results when using smaller round sizes, as we are giving more information back to the classifier. What we see in the graph above is just the opposite we would expect. We are getting better results when using bigger round sizes. That means that if we increase the number of queries in each search to do the search of relevant images, we are introducing noise instead of accuracy.

As we did with the query expansion strategy, we ran exactly the same experiment for the ASVM but using a different scope to ease studying the results.

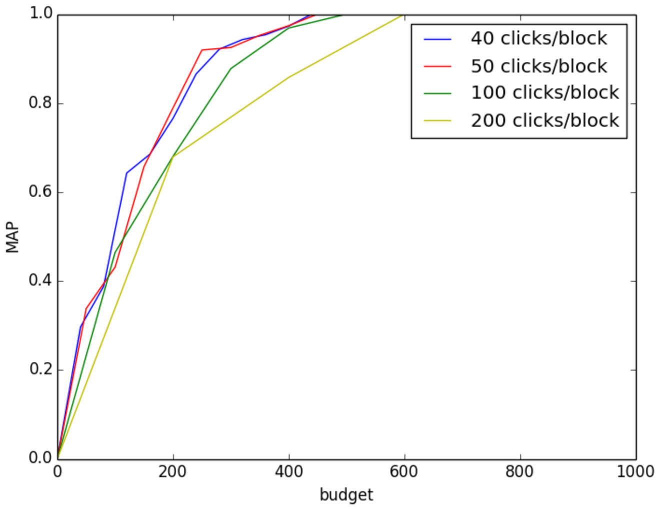


Figure 10. Different block sizes in the ASVM strategy.

So in the Figure 10 we used four different round size configurations to annotate the database using a scope between 40clicks/round to 200cliks /round. In this case, as we would expect and in difference to the query expansion strategy, when we are adding more iterations and implicitly we are giving more information back to the algorithm, we are increasing the accuracy. In all of these experiments we can see that at in the first iteration the MAP decreases. The reason why it happens is because in the first iteration, we do not have any previous query information to fit the classifier.

As we can see, the line where we are using 40cliks/round, we are getting better result than if we just do five iterations annotating 200images/round. Then we know that we are introducing more accuracy if using relevance feedback.

We also wanted to compare directly QEX versus ASVM for each relevant round size:

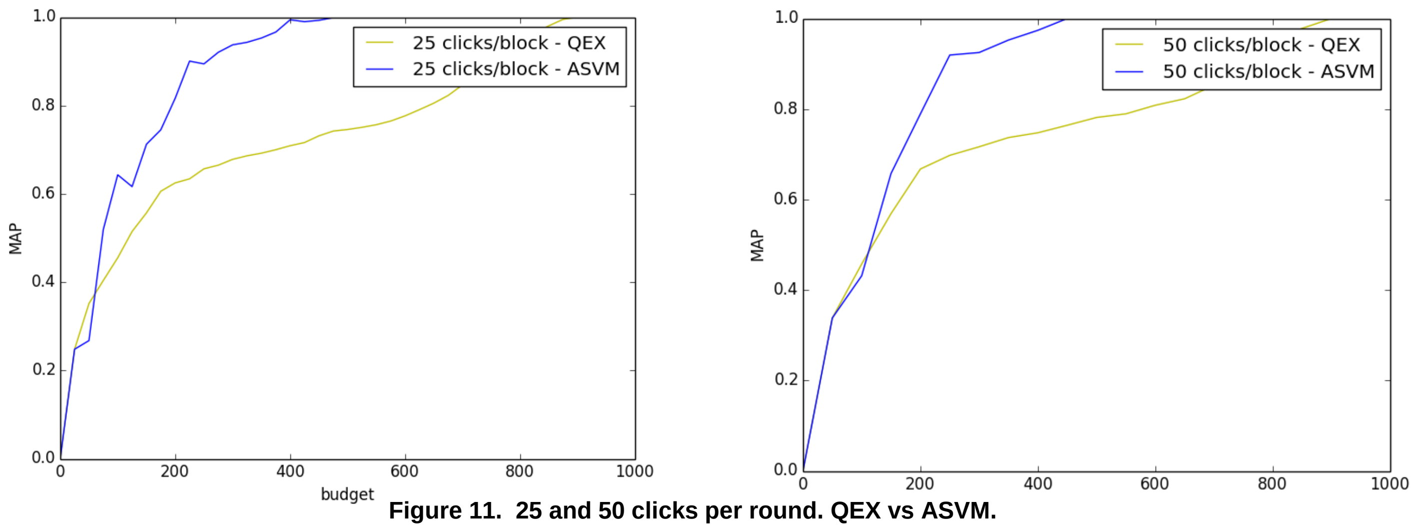


Figure 11. 25 and 50 clicks per round. QEX vs ASVM.

As we can see in the Figure 11, when comparing the two different relevance feedback strategies we are getting faster to the maximum MAP value (1.0) when we use ASVM rather than QEX.

So in this first experiments we saw that we have to use ASVM to add relevance feedback improvements. So in the next experiments, we focused on this strategy and then we tried to find the best distribution to get close to the EEG context.

4.3. <u>Optimal Round Size</u>

Once we know that we want to use the ASVM classifier to do our experiments from this point on, we want to find what would be the best round size configuration to reach faster the maximum MAP value. So we computed a different experiment where the scope was the round size, and each different line referred to different budgets spending: from 10% to

40% of the total amount of available clicks, where 100% corresponds to annotation the whole dataset of 1,000 images.

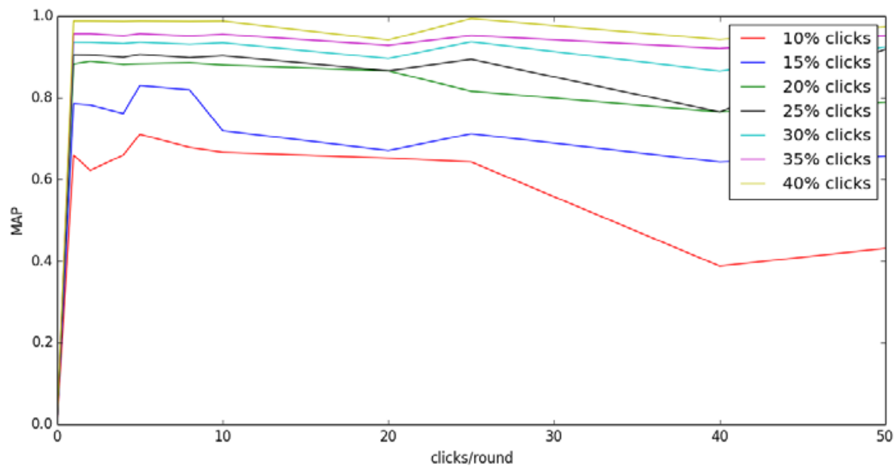


Figure 12. Different lines for different budget spending and their evolution in increasing round sizes.

It is interesting to see (Figure 12) that the maximum almost in all lines, is located around the round size of 25 clicks round. There are other local maximums with smaller round sizes, but we have also to consider that in computing and time spending terms is much easier to compute and train the classifier with bigger round as we will do less iterations.

So we can get two main conclusions from this graph:

1- In terms of budget, once we have annotated more than 20% of our database per query, we would get almost the same results using any round size between 1 and 50 clicks per round.
2- There is a tendency of local maximum at the round size of 25 that is interesting in terms of accuracy and computational ease.

4.4. <u>Distribution of Relevant Frames within a RSVP Block</u>

Here is the part where we try to match the previous results with the EEG requirements. As we said, to study the configuration that we should use in the EEG display, we have to consider that we cannot annotate two targets two close and also which percentage of relevant images we should include in each round.

We already know that we have to discard the best distribution that we have a priori. That would be annotate always images from the top of the scores list retrieved from the classifier. But we included this one in our studies just to compare it with rest and to have an optimum reference.

As we explained before, we purposed four different distribution strategies to see how the classifier behaves when using different annotation systems. The four different strategies are: Highest Scores, Random, 5-95%(10/190 in the graphs) and Clustering.

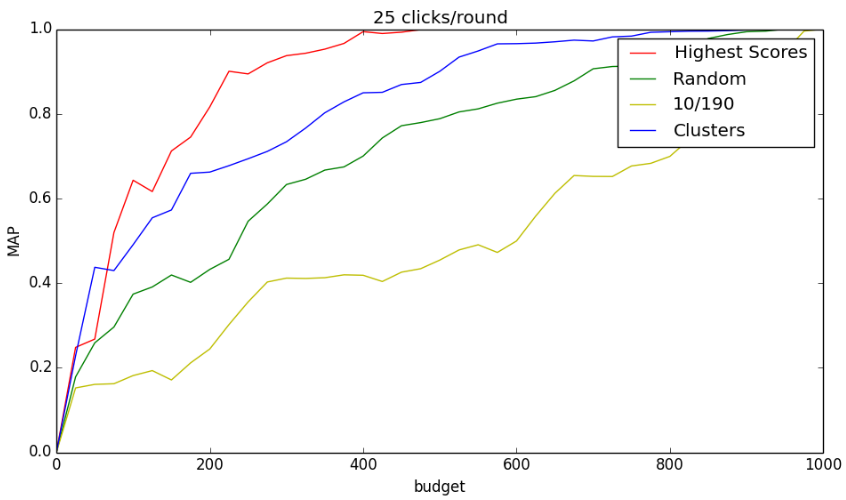


Figure 13. Different distribution strategies to annotate the images.

In the results in the Figure 13, the Highest Scores line has the best performance, as it is annotating always the frames with best scores. The problem with that strategy is that we cannot use it in an EEG context as it would display the targets too close in time, making it more difficult to generate and capture the P300 signal expected from relevant frames.

The “10/190” line refers to our 5-95% distribution as it is the same relation, and tries to state what would be the effect of relevance feedback if we would use the current display strategy that we have on the interface. As we can see in the plot it is not taking advantage of the information of the previous rounds as it is always fitting the classifier with images with the lowest scores and just a few relevant ones.

The most important lines are the “Random” and the “Clusters”. It is really interesting that both are taking advantage of the relevance feedback even they are not annotating all the relevant images at the beginning. That happens because of the diversity. In the random case, it is getting frames diversity because the images that it is annotating in each round are retrieved completely random.

We want to focus in the clustering strategy. As it is splitting the database in clusters for each round, the classifier is being fit with the maximum diversity as possible. Because of that, we can see that it overcomes the random strategy and it gets close to the highest scores strategy. This strategy would be possible to apply in an EEG context as it is labelling just a few relevant images per round.

Summarizing, we have a classifier that returns us a scored list of the whole database from where we have to pick a set of images to be annotated. So we explored some different ways to select the frames: Highest Scores, Random, 5-95% and Clustering. Each of them chose the images to be annotated following different strategies.

First we focused on the MAP performance of these strategies. We saw that using the Highest Scores we got faster to high MAP values but we already knew that this distribution would not have a good performance when using the BCI as it takes most of the relevant frames at the beginning instead of dividing them along all the rounds.

We also studied the distribution that we already used in our BCI experiments with real users. But it turned out to be the one with worst MAP performance because during the first rounds it was not using enough relevant frames.

Finally we studied two more distributions, Random and Clusters. On both of them we tried to use the maximum diversity as possible in each round. In the Random one, as it is logical we did not use the scored list to select the set to be annotated, instead we annotated random sets. On the other hand, the clusters distribution gathered into different clusters the frame descriptors more similar and this way we were easing the classifier to predict the new frames more efficiently. The results we got were quite good and the clusters strategy had better MAP values than the random one.

4.5. Minimum Separation between Relevant Frames

4.5.1. RSVP @ 5Hz

When using EEG signals we have to consider some other features as the frequency display and the distribution of the relevant images. Considering that we are working with a 5Hz frequency display, it means that when using a round size of 25 images per round, it would last for 5 seconds. As it is explained in the EEG studies part, to detect a relevant image with the EEG interface, it has to be separated at least at 1 second from the other relevant images. So in our case we could have 5 relevant images per round to consider that we are being successful in the annotation. Or in other words, we would like to have less than 20% of relevant images in our round.

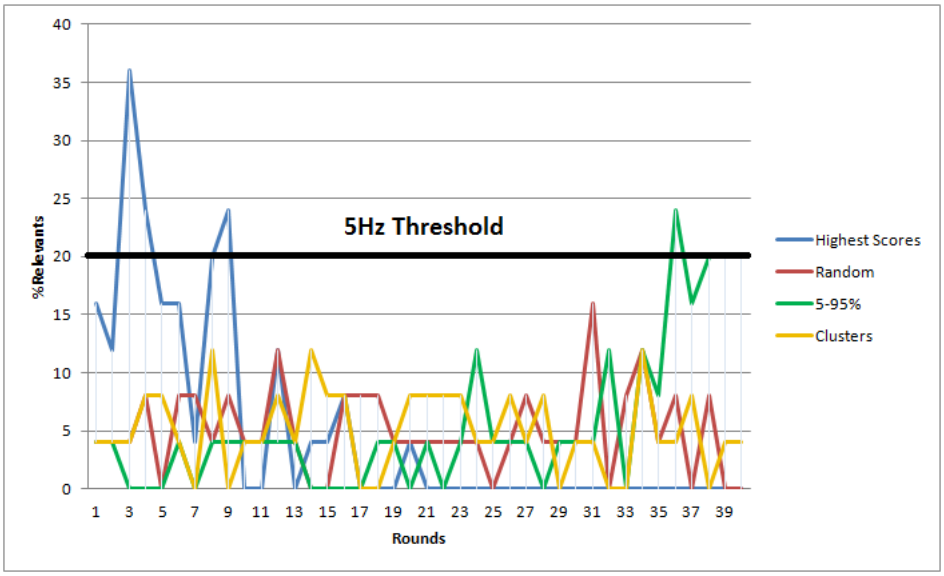


Figure 14. Percentage of relevant frames per round with the 5Hz threshold.

In next table we tried to state this features within the different strategies:

Table 1. Average percentage of relevant frames per round and success percentage for each distribution. Frequency at 5Hz.

Strategy	Highest Scores	Random	5-95%	Clusters
Relevants/round	14.29%	6.25%	7.14%	6.25%
Succes Percentage	71.42%	100%	89.28%	100%

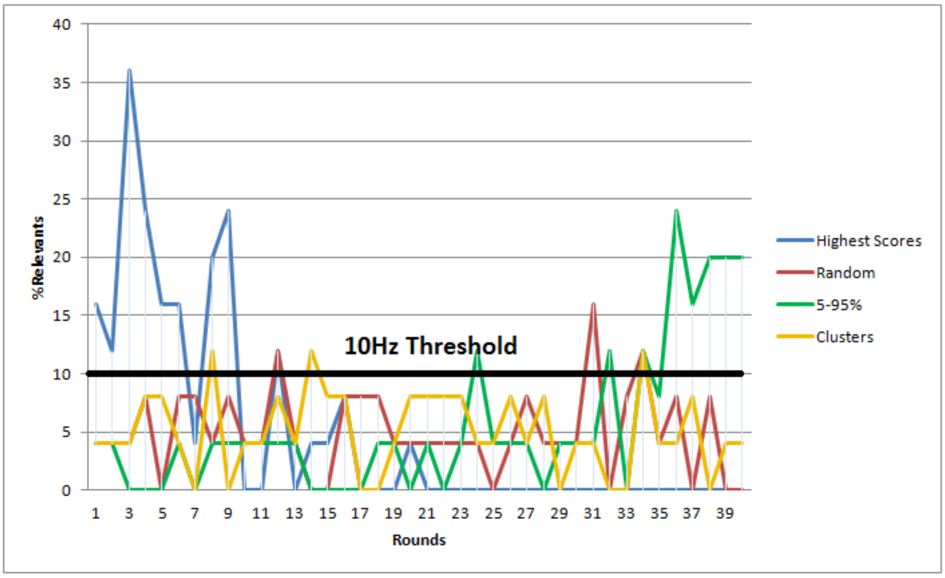


Figure 15. Percentage of relevant frames per round with the 10Hz threshold.

In the first row we are showing the average percentage of relevant images per round for each strategy. In this case, we are only considering the rounds where at least there is one relevant frame. We observe that only the Highest Scores strategy overpasses the limit.

Then in the second row we have the average percentage of success. We consider a successful round when we are below the limit. As we can see, we get good results when using the random and clustering strategies and worse results if we would use the highest scores strategy to annotate with the EEG interface.

4.5.2. **RSVP @ 10Hz**

To include more results in our study, we have also considered that we worked at 10Hz. At that frequency each round of 25 images would last just for 2.5 seconds and that means that we could just have between 1 and 2 relevant images. It is the same than saying that we should have less than 10% of relevant images in each round to be able to label them using EEG signals

Table 2. Average percentage of relevant frames per round and success percentage for each distribution. Frequency at 10Hz.

Strategy	Highest Scores	Random	5-95%	Clusters
Relevants/round	14.29%	6.25%	7.14%	6.25%
Succes Percentage	35.71%	90.62%	71.42%	90.62%

Here we can see, if working with 10Hz rate, there are less strategies that we could use to annotate the images with the BCI. As we can see in the graph above, the 5-95% strategy would be out of the boundaries and we could just use the Random and Clusters strategies.

To be consistent with the experiments already carried out using the BCI, we wanted to study some other strategies beside the 5-95% distribution. We wanted to try with different relations to see if we could avoid having most of the relevant images at the last rounds, and this way try to have a more flat curve in the last graph. So we run the experiments again using 10-90% and 20-80%. Just to remember, that means that for example in the second configuration, we would pick 20% of the images to be annotated from the top of the scores list, and 80% from the bottom of the list.

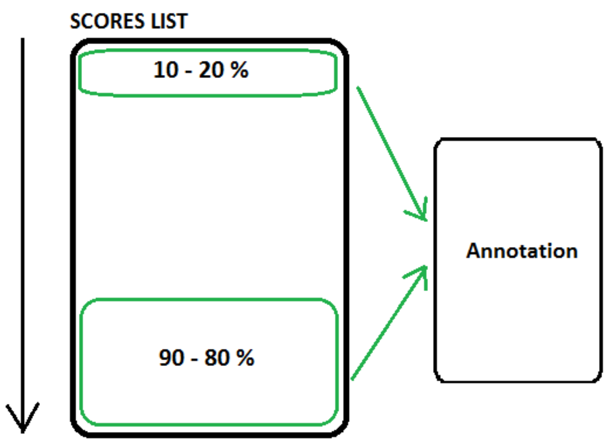


Figure 16. Small part retrieved from the top of the scores list and main part from the bottom but with different relations.

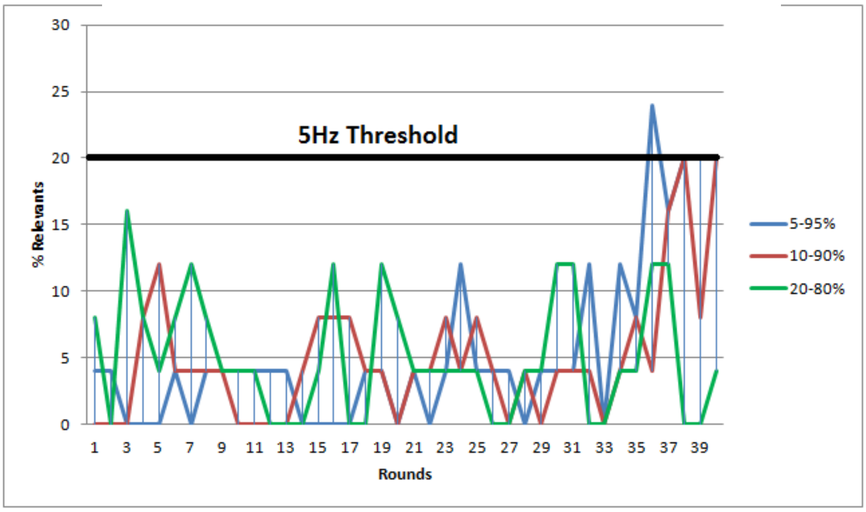


Figure 17. New purposes of distribution with the 5Hz threshold.

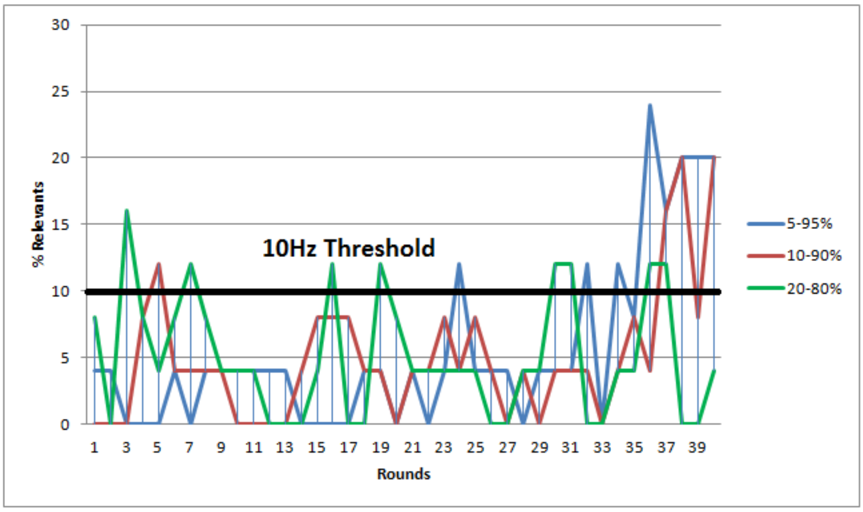


Figure 18. New purposes of distribution with the 10Hz threshold.

If we focus on the figures 17 and 18, we can see that if we use different percentages we can improve the curve shape. Using the 5Hz frequency, we would be under the threshold almost in all of the configurations so we could use any of them at this frequency. But if we look to the 10Hz threshold, we can see that we are over the line in all of them. So if we understand the 10Hz line as the success threshold, then we should choose the flattest curve and the one that keeps more time under the threshold.

Table 3. Average percentage of relevant frames per round and success percentage for each distribution.

Strategy	5/95	10/90	20/80
Relevants/round	7.85%	6.89%	7.14%
Succes Percentage (5Hz)	89.28%	93.10%	100%
Succes Percentage (10Hz)	71.42%	86.20%	72.41%

As we can see in the Table 3, almost all of the configurations have the same average percentage of relevant frames per round. So if we focus the number of sample that are under the success boundaries, we can see that for the 5Hz frequency choosing 20% of frames from the top of the scores list and 80% from the bottom, we have 100% of success percentage. That percentage means that in all the rounds the percentage of relevant images was under the threshold that allows us to annotate them with the BCI. At the same time, if we switch the frequency to 10Hz, then the best configuration would be 10 - 90%.

5. Conclusions and future development:

In this thesis we wanted to find a way to include Relevance Feedback techniques in an image retrieval system using an EEG-based brain-computer interface and a RSVP display.. So first we studied the different classifier that we could use, whether Query Expansion (QEX) or Support Vector Machine (ASVM).

We did some experiments using the Mean Average Precision to evaluate the results and it turn out to be that the Support Vector Machine classifier worked much better. In fact the Query Expansion strategy was just introducing noise to our search. After this conclusion, QEX was discarded and we started working just with the ASVM.

So from this point on we wanted to know what size should have our iteration rounds in order to get the best performance. To decide that we considered two different points:

1- The longer were the round sizes, the easier to compute the distances by the classifier.
2- Find the longest round size with the maximum MAP value.

After running the experiment with different budget expending and for different round sizes we realized that for all the different budgets, there was a tendency to decrease the MAP value after the 25 images per round. So we choose 25 images per round as our optimum theoretical value.

All this study must be related with the EEG annotation systems, so we tried to find a way to related them efficiently. We had to know that the BCI had some constraints due to the display frequency and the brain signals responses. We can just detect one image per second using the brain information because as we explained the P300 signal last around 1 second after the target image is displayed. Mostly we worked at 5Hz, it means that we had to find a way to annotate at most 5 relevant images per round.

After knowing the MAP performance for the different strategies, we wanted to know if it was possible to apply them on the BCI. So we run the experiments again focusing on the number of relevant frames there were per iteration and for each different distribution.

We knew that in order to label them using the EEG signals, due to its frequency and time response constraints we had limit of relevant images per round percentage. Using the frequency display of 5Hz we couldn't overcome the 20% and with the 10Hz frequency the limit was at 10%.

When using the 5Hz frequency, we realized that we could use all the configurations almost without looses. But some of them had much better performance. As we said, it depended on how good it distributed the relevant frames along the iteration rounds so we had much better results in the Random and the Clusters strategies.

In the 10Hz case, it would not be possible to use the Higher Scores and the 5-95% as in the first case it would only be possible to annotate 35% of the rounds properly and in the second distribution we could only annotate a 71.4% of the rounds.

We found interesting to add other studies in the 5-95% distribution as it is the one already used in the EEG experiments. As we explained the problem in this case was that most of the relevant frames were placed at the last rounds. So we tried to spread the scope and did the same study but with 10% frames from the top of the scores list and 90% with low scores, and the same with the 20-80% relation.

We saw that at both frequencies the distribution used in previous BCI experiments in our lab was actually the worst. If we spread the part we get with high scores, we see that we can divide more efficiently the relevant frames thorough all the iterations. So if we want to run the EEG experiments at 5Hz, then we would recommend using the 20-80% frames distribution, as it would still be getting 100% of success possibilities in each round. Otherwise, if we wanted to run the experiments at 10Hz, then we should switch the relation to 10-90% as it shows the best performance.

After all this study, I think that now it would be possible to implement an EEG annotation including the Relevance Feedback feature. Once known the constraints of the EEG annotation, we would just need to focus on the points explained in that thesis to choose the best configuration.

I think it could be really helpful in those cases where is needed to annotate huge databases. With the studies and predictions did in that thesis, it could ease to know what budget of the database we should label, which classifier we should use and also what kind of distribution of the data we should set. This way it would not be needed to annotate the whole database and it would save time and resources.

Finally, I would also like to highlight that my collaboration on the parallel work with my co-advisors to compare mouse vs EEG interfaces has resulted in an 8-page scientific paper to the ACM International Conference on Multimedia Retrieval (ICMR). I am the third author of that paper, out of a total of seven. The review of that paper is not available yet at the time of submitting this thesis report for evaluation and, for this reason, it has not been included as an annex to this thesis report.

5.1. Future Work

I consider that there are two points that would be really interesting to do more research on them:

1- In the clusters distribution to decide which images must be annotated, we always chose from each cluster the frame descriptor closest to the centroid. Taking into account that we have the ground-truth file to label the images with our software, it could be interesting to add that information when retrieving the images from the clusters. Then we could see if it improves the MAP values and if it gets closer to the Highest Scores strategy.

2- During all our experiments, we have been using binary labels to know whether the frames were relevant or non-relevant. I think it could be interesting to use label that simulate the signals that we get from the EEG signals. These values are not binary and it is possible to simulate them using regressors. It would add a more realistic situation to all the experiments.

Bibliography:

[1] Eva Mohedano Graham Healy, Xavier Giró-i-Nieto, Kevin McGuiness, Noel O'Connor, Alan F.Smeaton. "Object Segmentation in images using EEG signals". Dublin City University.

[2] Simon Tong, Edward Chang. "Support Vector Machine Active Learning for Image Retrieval". *Standford University.*

[3] Aida Rubiano Márquez. " Visual Search with Relevance Feedback based on Weights Updating". *Universitat Politècnica de Catalunya.* http://bitsearch.blogspot.com.es/2010/03/relevance-feedback-for-image-retrieval.html?q=relevance+feedback

[4] Jun Wang, Eric Pohlmeyer, Barbara Hanna, Yu-Gang Jiang, Paul Sajda, Shiih-Fu Chang. "Brain State Decoding for Image Retrieval". 1Department of Electrical Engineering, Columbia University, New York, NY 10027, USA.

[5] Li Jing, Liu Xianzeng, Ouyang Gaoxiang. "Using Relevance Feedback to distinguish the Changes in EEG During Different Absence Seizure Phases". Clinical EEG and neuroscience : official journal of the EEG and Clinical Neuroscience Society (ENCS), 2014; doi:10.1177/1550059414548721.

[6] D. Fernandez-Canellas. "Modeling the temporal dependency of brain responses to rapidly presented stimuli in erp based BCI". Master's thesis, Northeastern University, 2013.

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

[8] A. Kapoor, P. Shenoy, and D. Tan. Combining brain computer interfaces with vision for object categorization. In Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2008.

[9] I. Pathirage, K. Khokar, E. Klay, R. Alqasemi, and R. Dubey. A vision based p300 brain computer interface for grasping using a wheelchair-mounted robotic arm. In 2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pages 188–193, July 2013.

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

[11] Amaia Salvador. "Exploiting User Interaction and Object Candidates for Instance Retrieval and Object Segmentation". Universitat Politècnica de Catalunya. https://imatge.upc.edu/web/sites/default/files/pub/xSalvador_0.pdf

[12] Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *Computer Vision–ECCV 2014* (pp. 584-599). Springer International Publishing.

[13] Yangging Jia, Evan Selhamer, Jeff Donahue, Sergei Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, Trevol Darrel. ."Convularional Architecture for Fast Feature Embedding". UC Berkeley

Appendices:

- Workpackages description.
- Python scripts and code used to run the experiments.

6. Annex 1: Work Packages

The first idea of this thesis was to annotate images using brain signals and then compare the results with the mouse interface. After collecting some data from some real users and checking that it was actually possible to annotate images with BCI, we found more interesting to point the project to the Relevance Feedback part and then try to relate it with the first idea. The comparative study of BCI vs Mouse interface was run in parallel by my co-advisors and, while I participated in the study, that was not the main task related to my bachelor thesis.

So we did some changes in the last packages, as at beginning we wanted to study the relevance feedback techniques directly from the results got from the BCI experiments. But finally we explored the relevance feedback beside the EEG results carried out mainly by my advisors.

<table>
<tr><td>Project: Instance Search using EEG Signals</td><td colspan="2">WP ref: (WP#) 1</td></tr>
<tr><td>Major constituent: Background</td><td colspan="2"></td></tr>
<tr><td rowspan="2">Short description: Study of the functioning of the EEG hardware system. Study the results of the previous research in this field. Study how to get and how to manage large image databases.</td><td colspan="2">Planned start date: 22/09/2014
Planned end date: 13/10/2014</td></tr>
<tr><td colspan="2">Start event:22/09/2014
End event: 10/10/2014</td></tr>
<tr><td>Internal task T1: First steps in the required computing language and interpretation of the previous EEG systems results in that context.</td><td>Deliverables:</td><td>Dates:
10/10/2014</td></tr>
</table>

<table>
<tr><td>Project: Instance Search using EEG Signals</td><td colspan="2">WP ref: (WP#) 2</td></tr>
<tr><td>Major constituent: Experiment Set Up</td><td colspan="2"></td></tr>
<tr><td rowspan="2">Short description: Preparing all the tools that will be needed for the first test.</td><td colspan="2">Planned start date: 10/10/2014
Planned end date: 20/10/2014</td></tr>
<tr><td colspan="2">Start event: 10/10/2014
End event: 30/10/2014</td></tr>
<tr><td>Internal task T1: Ask for the access at the BCI lab and for the needed material.</td><td>Deliverables:</td><td>Dates:</td></tr>
</table>

Project: Instance Search using EEG Signals	WP ref: (WP#) 3	
Major constituent: Simulation and Processing		
Short description: First test of the EEG hardware system to a single user. Set up of the software taking into account the impulses of the test. This stage will ensure the proper functioning of all the elements for the final user brain images retrieval.	Planned start date: 21/10/2014 Planned end date: 7/11/2014 Start event: 21/10/2014 End event: 7/11/2014	
Internal task T1: Test in the laboratory with a single user. Internal task T2: Processing software for the image retrieval process.	Deliverables:	Dates:

Project: Instance Search using EEG Signals	WP ref: (WP#) 4	
Major constituent: Data Acquisition		
Short description: Data acquisition from the different brain users in the lab and process of this data.	Planned start date:10/11/2014 Planned end date:14/11/2014 Start event: 10/11/2014 End event:16/11/2014	
Internal task T1: Find the candidates to do the experiment and arrange the trial meetings.	Deliverables:	Dates:

Project: Instance Search using EEG Signals	WP ref: (WP#) 5	
Major constituent: Result Analysis		
Short description: State the differences between both annotation algorithms and take conclusions.	Planned start date:12/11/2014 Planned end date:26/11/2014 Start event:12/11/2014 End event:15/11/2014	
	Deliverables:	Dates:

<table>
<tr><td>Project: Instance Search using EEG Signals</td><td colspan="2">WP ref: (WP#) 6</td></tr>
<tr><td>Major constituent: Relevance Feedback</td><td colspan="2"></td></tr>
<tr><td rowspan="2">Short description: Finding the best classifier to use the relevance feedback techniques.</td><td colspan="2">Planned start date: 15/11/2014
Planned end date: 21/12/2014</td></tr>
<tr><td colspan="2">Start event:15/11/2014
End event:30/12/2014</td></tr>
<tr><td>Internal task T1: Query expansion experiments.
Internal Task T2: Active Support Vector Machine experiments.
.</td><td>Deliverables:</td><td>Dates:</td></tr>
</table>

<table>
<tr><td>Project: Instance Search using EEG Signals</td><td colspan="2">WP ref: (WP#) 7</td></tr>
<tr><td>Major constituent: Composition of relevant and non-relevant frames.</td><td colspan="2"></td></tr>
<tr><td rowspan="2">Short description: Explore the different composition to try to find the one that fits better with the BCI constraints.</td><td colspan="2">Planned start date: 20/11/2014
Planned end date: 30/12/2014</td></tr>
<tr><td colspan="2">Start event:20/12/2014
End event:10/01/2015</td></tr>
<tr><td>Internal task T1: Experiments with the different compositions.</td><td>Deliverables:</td><td>Dates:</td></tr>
</table>

Project: Instance Search using EEG Signals	WP ref: (WP#) 8	
Major Constituent: BCI composition analysis with Relevance Feedback.		
Short Description: Defining quality measures for the different compositions.	Planned start date: 10/12/2014 Planned end date: 21/12/2014	
	Start event:5/01/2015 End event:15/01/2015	
Internal task T1: Studying the results.	Deliverables:	Dates:

Project: Instance Search using EEG Signals	WP ref: (WP#) 9	
Major constituent: BSc thesis development.		
Short description: Deliverables of the documentation required and preparation of the oral defense.	Planned start date: 22/09/2014 Planned end date: 21/01/2014	
	Start event:22/09/2015 End event:18/02/2015	
Internal task T1: Project Proposal and Workplan delivery. Internal task T2: Critical Review delivery. Internal task T3: Final Report delivery. Internal task T4: Oral Defense Preparation	Deliverables:	Dates: