

IMPRESSION EXTRACTION FOR BUILDINGS BY USING EEG

DEPARTMENT OF SYSTEM DESIGN ENGINEERING

Keio University



STUDENT: MARIO COROMINAS

ADVISER: YASUE MITSUKURA

JULY 2013

Contents

1	Introduction.....	2
2	Background.....	3
2.1	Introduction to EEG	3
2.1.1	EEG Generation	4
2.1.2	Brain Rhythms	4
2.2	MITSUKURA LAB’S background.....	5
3	Methods used.....	6
3.1	PCA – Principal component analysis	6
3.2	CV – Cross Validation.....	7
3.3	LDA – Linear Discriminant Analysis	8
3.4	SVM – Support Vector Machine.....	8
3.5	BE – Backward Elimination	9
4	Experiment	10
4.1	Definition.....	10
4.1.1	Study indoor or outdoor.....	11
4.1.2	Study buildings from different places.....	12
4.1.3	Study real scenarios, pictures with/without manipulation	13
4.1.4	Summary of final features	14
4.2	Preparation.....	16
4.2.1	Eight manipulated pictures.....	16
4.2.2	Program in MATLAB	18
4.2.3	Questionnaire.....	19
4.3	Experimental development	20

4.3.1	Procedure of the experiment	21
4.3.2	Extracted data cleansing.....	22
4.4	Analysis and results	22
4.4.1	Noise detection	23
4.4.2	File organization	24
4.4.3	Grouping	27
4.4.4	Initial conditions – variables set up and filter diagram sequence	28
4.4.5	First classification method.....	30
4.4.6	Combination of frequencies from the first classification method	39
4.4.7	Second classification method	41
5	Discussion	44
6	Conclusion	45
7	Bibliography.....	46

1 Introduction

Nowadays, in the architecture there are a lot of subjective opinions about how beautiful or likeable they are. Only the judgment of respected people could be taken into account. However, to do this project we believe there could be a way to make a difference in this field.

If there was any way to get an objective impression from buildings it would be a powerful tool in the architecture field. The objective impression from people should be directly extracted from their thoughts. That could sound like science fiction, but today it is possible in some specialized laboratories. The MITSUKURA LAB is specialized with the EEG non intrusive devices that measure a simple voltage difference from the frontal lobule part of the brain and the ear.

From the frontal lobule it's possible to detect the LIKE and DISLIKE feelings from people. So it was believed that doing the right experiments and analysis afterwards it could be a starting point for that previous idea commented in the last 2 paragraphs.

➤ Objectives

The objectives of this project would be the following ones:

- 1) Determine one's preferences (like/dislike) on buildings by reading their brain waves.
- 2) Analyze different visual aspects on buildings that might be detected by reading the brain waves.

2 Background

To understand the global importance of this project it should be necessary to introduce some of the concepts related with this study. It's also important to understand the background of the MITSUKURA LAB not only for this project but also because it will give a deeper and wider vision of this field.

2.1 Introduction to EEG

The concept of EEG started on 1875 when Richard Caton measured for the first time the brain activity in the form of electrical signals. The meaning of EEG would be Electro- (referring of the registration of electrical signals) Encephalo- (referring to emitting the signals from the head) and Gram (referring to drawing or writing).

At the beginning the main purpose of studying and understanding the brain waves were for medical reasons. The epileptic seizure was one of those first targets to be focused by the study of EEG signals but after the 1950s they started to expand the investigations in other brain illnesses.

From then, the study of neuronal activity has increased in many other fields and not only epileptic seizure but also for loss of conscience, behavioral disorder, sleep disorder or just to analyze the electrical activity in the brain.

2.1.1 EEG Generation

An EEG is a current that flows during the synaptic excitations of the dendrites of many pyramidal neurons in the cerebral cortex. When the neurons are activated, a current goes through the dendrites and that current generates a magnetic field that can be measured.

Our heads have some thick layers like the brain, the scalp and the skull; also there are some other thinner ones in between. There is a lot of noise created in the brain and outside external layers so it's necessary to have a large number of neurons generating signal to get a recordable signal. Afterwards, those signals are greatly amplified.

2.1.2 Brain Rhythms

There are brain rhythms on those electrical waves and it has already been tested and confirmed that there exist differences between wakefulness and sleep. The amplitudes and frequencies change from a human to another one. Those characteristics also change with the age.

So the main brain frequencies that we could find are:

- a) Delta (δ) \rightarrow 0.5-4 Hz
- b) Theta (θ) \rightarrow 4-8 Hz
- c) Alpha (α) \rightarrow 8-13 Hz
- d) Beta (β) \rightarrow 13-30 Hz
- e) Gamma (γ) \rightarrow 30-45 Hz

2.2 MITSUKURA LAB'S background

This lab is pioneer on this kind of researches. They always work with brain preferences, extracted from KANSEI (acquired after birth) and not with sensitivity (that would be acquired as an innate ability). Those preferences produce some Bio-signal waves that are possible to be detected by EEG.

The biological signal permits to obtain a sequential measurement for each second and permits to avoid questionnaires. If the brain signal is correctly interpreted, the EEG is more reliable than a questionnaire.

The EEG signals might be different varying on the people who are evaluated. Then it's needed to have a easily wearing device like the one used in this project (a Hair-band type device) that doesn't need gel and it can be used in anyplace at anytime. This device permits to measure a simple signal without stressing the participant and it's particularly useful to do multiple experiments at different times (and for short duration).



MITSUKURA LAB's EEG device

There is also needed a strict signal processing and a signal processing that takes into consideration individual differences.

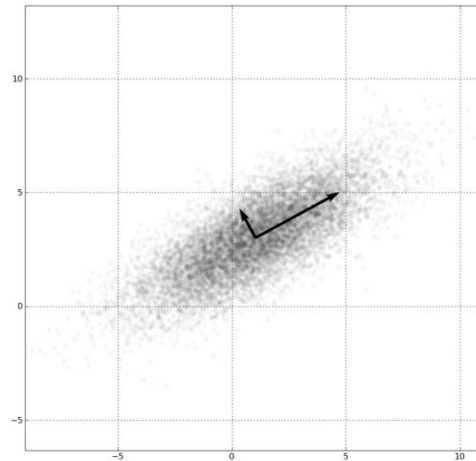
The application of MITSUKURA LAB researches are some as Acoustic sense recognition (preferences on music), Taste sense recognition, Haptic recognition, Visual recognition & acoustic recognition (degree of interest in TV commercial evaluation, stress degree evaluation, sleepiness degree evaluation).

3 Methods used

In the experiment chapter there are some methods used that should be introduced before. Here it will be a very short introduction to all of them so the reader can get familiarized with them.

3.1 PCA – Principal component analysis

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called 1st PCA), the second greatest variance on the second coordinate, and so on.



PCA of a scatter plot. The vectors shown are actually eigenvectors of the covariance matrix

This method converts a set of observations that might be correlated into a set of linearly uncorrelated variables. It's a very useful tool in signal processing and almost in all researches is needed.

3.2 CV – Cross Validation

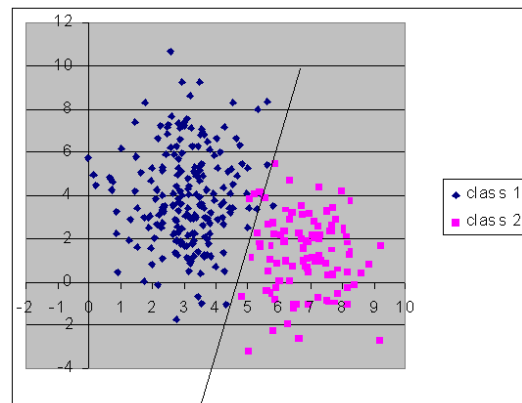
This method is a practical technique to repeat a set of experiments and give a reliable result. The basis is to separate the whole data participating on the experiment in small even groups. When the experiment is ongoing, there should be as many repetitions as groups. In each repetition there is a different group as a testing data and the rest of them as training data.

Usually, in statistics, the whole amount of data is divided by 20 different groups so each repetition there is a 5% of data treated as testing and 95% as training.

3.3 LDA – Linear Discriminant Analysis

The LDA has a lot of possible utilizations. One of them would be the linear classifier and that is what it's going to be used in this project. The LDA is closely related to PCA because they both look for linear combinations of variables which best explain the data. However, the main difference between the LDA and PCA is that the first one attempts to model the difference between the classes of data.

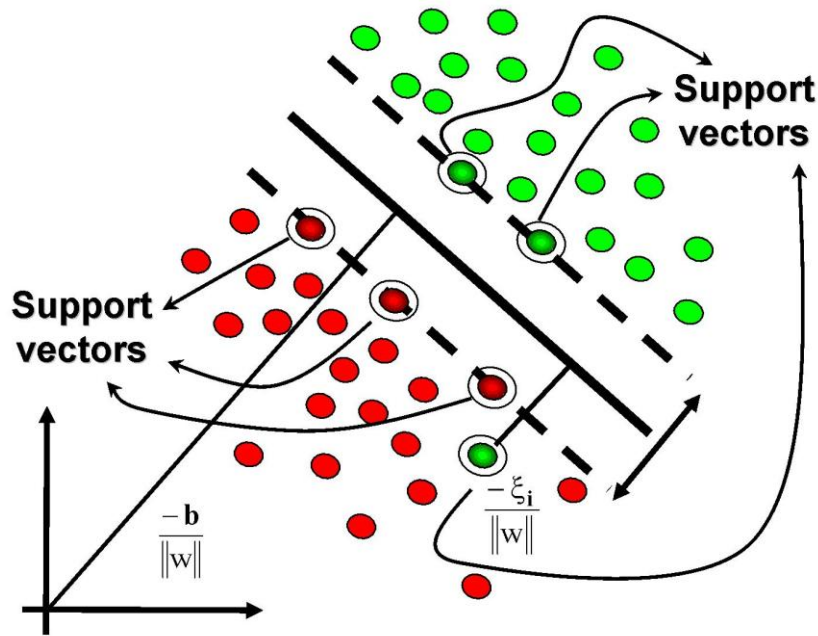
The MATLAB function would be `classify` which also allows using a quadratic order function. So both options will be used in this project.



LDA linear classification example

3.4 SVM – Support Vector Machine

This method is another kind of classification data tool. It can be considered as alternative method from LDA. The SVM provides a classification non-linear that uses hyper planes. The SVM defines a function in between the different kinds of data by using as a support vectors defined thanks to some data.



SVM graphic draft

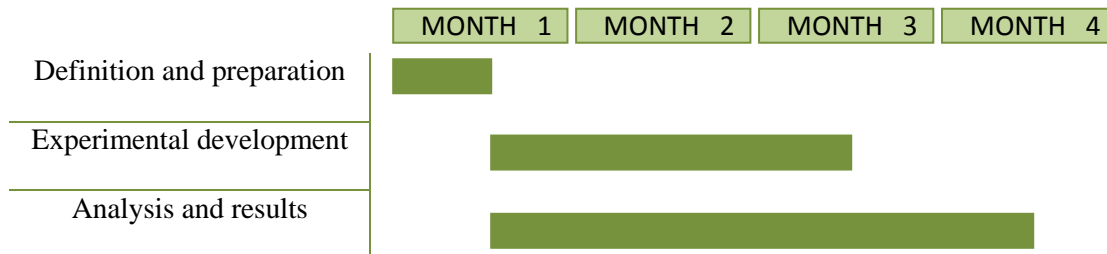
3.5 BE - Backward Elimination

This method is very useful when there a lot of variables participating in a experiment with an output result. The intention is to discover a good combination between all those variables to get the best result as possible.

So if this method has N variables, it will have N-1 steps. On the first step the experiment will be repeated N times and it will remove a different variable each time. At the end it compares the results and removes the variable which gives the combination with the best result. This will be repeated in each step (every time with -1 variable) and finally a comparison has to be done between the best results from each step and keep the best one. Then it's possible to check the records to see which variables did the combination that gave the best result.

4 Experiment

This chapter of the project has 3 different parts. The first part took 2 weeks and it could be named as definition and preparation. The second took 2 months and could be called as the experimental development. Finally, the third and last part took 3 months but the first 2 months of that last part were superposed to the previous part, this part would be the analysis and results.



Then, all those 3 parts will be explained in this chapter and will be carefully described. There was not too much time in this project, so the 2nd and 3rd parts had to be done together. However, if there were more time it would have been better to split *Analysis and results* in 2 different parts and also add a new part named as *program development*.

4.1 Definition

The first part would have two different steps. The first one would be to define the conditions and the second one to obtain and prepare the experiment.

So first of all, in this project it was needed to define the parameters to study from buildings. To do so it was considered different possibilities: study indoor or outdoor, study real scenarios or pictures of real buildings or manipulated pictures, study buildings from different countries/cities or from the same city.

4.1.1 Study indoor or outdoor

When we want to analyze a building we need to look for its outdoor sight and for its indoor sight.

To decide which one to look at we looked for both convenient and inconvenient issues.

The outdoor sight provides the first impression on a building. There is no need to look at it for long time, in just a few seconds you can realize if you like it or not. In addition it's easier to do the experiments even if they are in pictures or in reality.

The indoor sight provides you the feeling of living in that building because it would be surrounding you so it would be easier to stimulate the senses of your body and then extract the significant waves from the brain. The feeling that you can get from the inside is much deeper than the one that you could get from outside and it's easy to believe that there could exist more differences on the brain waves than looking just from outside.

So finally the decision was to do it from outside for various reasons. It is true that the indoor could provide more differences that would make easier to classify the data afterwards; however, there are plenty of inconvenient if choosing that option.

- a. To obtain the data from the indoor sight it would be needed a lot of time (long time experiments) to make sure that the person doing the experiment feels involved in the building.
- b. Long time experiments are really hard to manage and to prepare the living areas. The time for this project was 4 months.
- c. If the experiment was decided to be done by pictures (all with short duration) the outdoor one is more reliable to the reality than the indoor. Because a picture of the outside will be

very similar at how you would see it whereas the indoor picture wouldn't be able to surround you to make you feel like if it was real.



Outdoor picture. Easy to get an overall impression in few seconds



Indoor picture. It's hard to get involved in the whole house or room.

4.1.2 Study buildings from different places

This option was proposed actually by one architect from Keio, Jorge Almazán. His idea was to use different atmospheres to make even wider differences on the signal waves. In this case, the problem also was the time and the budget (too much budget for the small time of this project). Anyway, it's interesting to consider this option just in case there is a new project coming, bigger and more complete than this one.

In this project just one city or atmosphere would take place, so this option cannot be carried out.

In the followings paragraphs it will be decided how many places and the type of them.

4.1.3 Study real scenarios, pictures with/without manipulation

This decision was very important for the posterior study because it would give some initial conditions that might change other aspects from the project. Each option has some advantages on the others.

A real scenario would probably be the best one to get the most realistic impression and consequently get higher probability to succeed on classifying correctly between like and dislike waves. However, this option had to be discarded due to the difficulties to perform the experiments in real scenarios. Internet connection would be needed to run MATLAB, it would be needed a lot of time to get into the scenarios and the noise and conditions would be very difficult to control. So the noise and smell would probably affect on the person doing the experiment and it would be difficult to analyze the difference between the sensitivity of smell or noise and the one from the sight. This option was discarded for those reasons.

Using real pictures of buildings could be a good option because the manipulation of them could probably give the feeling of artificial to the person doing the experiment. But the problem of using real pictures is that it would be very hard to define the parameters of the picture. Those parameters could be different adjectives or variables easy to determine in between the pictures that might be used to study.

Manipulated pictures ended up being the best option due to the flexibility of it. It's possible to use the same background and just change the variables or parameters that are required. The best reason for choosing this option was that while using the same background in all the pictures you make sure that all the contour conditions on the picture are fixed as desired.

4.1.4 Summary of final features

After looking at those parameters and deciding which will be the initial conditions it seems to be that the experiment will be done within the following aspects:

- ✓ It will be an outdoor scenario. Just one picture from a building that will be modified to obtain new different pictures that will be evaluated in the experiment.

At this point it was necessary to define which aspects of the picture would be manipulated. To decide the main aspects to study, it was an accord between the weekly meetings and the verification of the professor Almazán (Architect from Keio University, *Architecture studiolab*). Those variables have been selected due to their simplicity to visualize and easy to manipulate. So the main aspects to consider in the modifications were 3:

- 1) Green: one half of the pictures would be with a lot of green (trees and grass) around the house and the other half without it.
- 2) Color: one half of the pictures would be with a cool color and the other half with calm color
- 3) Windows: one half would be with a lot of windows and the other one without windows.

So the final combination between those 3 variables would be like the following table shows:

rows:	Color	Green	Windows
1	calm	yes	yes
2	calm	yes	no
3	calm	no	yes
4	calm	no	no
5	cool	yes	yes
6	cool	yes	no
7	cool	no	yes
8	cool	no	no

So that makes a total of 8 different pictures that are necessary to be manipulated. The idea is that all of them come from the same background so the differences between those pictures are mainly the ones defined in the table above. There was a painstaking research looking for a real building but easy to manipulate in the internet and finally it was chosen this one:



Original picture

4.2 Preparation






For the preparation for the experiment there were 3 different aspects to get ready. One would be to prepare the 8 pictures that were defined in the definition process; another one would be to prepare the program in MATLAB and the last one to prepare the questionnaire to be answered for each picture. So the following points of this report would be about those 3 issues that needed to be done.



4.2.1 Eight manipulated pictures

To do the manipulations on that original picture it was used the Photoshop CS6. A previous training had to be done to learn how to use the basic and more useful tools of Photoshop. After a week all the pictures were done.

Those were the results:

rows:	Color	Green	Windows	Result
ORIGINAL				
1	calm	yes	yes	

2	calm	yes	no	
3	calm	no	yes	
4	calm	no	no	
5	cool	yes	yes	
6	cool	yes	no	

7	cool	no	yes	
8	cool	no	no	

4.2.2 Program in MATLAB

The idea for the experiments was to show the 8 pictures in a random sequence to make sure that there couldn't be any correlation in between the order of the pictures and the like/dislike preferences. This was the main reason to decide to use MATLAB as the photo player.

To make the program it was important to take into account:

a. RANDOMLY

The pictures should be selected randomly and the program should make sure that only 8 pictures would be displayed without repeating any of them.

b. ROUTINE FOR EACH PICTURE

There should be 10 seconds of white screen before and after the 10 seconds of picture. Also a countdown could be included to let the person doing the experiment know that the next routine is coming.

c. MANUAL PAUSES

As the participant should be able to answer the questionnaire after each picture, a manual pause should exist. The program needs to be activated manually to start the next routine for each picture.

d. RECORD SEQUENCES

The program should be able to record the order of the sequence that the random function is creating. This is very important to pair the questionnaires and brain waves records with the corresponding picture.

4.2.3 Questionnaire

This should be done in the computer first and they should be printed afterwards. The idea was that each participant had the questionnaire in paper in front of them so they could easily answer them while watching the experiment on the computer’s screen.

The most important features to include in the questionnaire were already decided when choosing the variables to modify on the original picture. Some essential questions would be the ones related with the modified variables. So the questionnaire was the following one:

Semantic Differential Questionnaire

GREEN:	Natural	3	2	1	0	1	2	3	Artificial
COLOR:	Calm	3	2	1	0	1	2	3	Cool
WINDOWS:	Adequate	3	2	1	0	1	2	3	Inadequate
OVERALL:	Good	3	2	1	0	1	2	3	Bad
What are your feelings about the picture? (please, answer freely)									

As it is possible to see in the figure above is there are 2 different kinds of ratings.

The 1st one is numerical and it has 4 different fields (green, color, windows and overall). This kind of questionnaire is called Semantic Differential Questionnaire (SDQ) and it's commonly used for this purpose. There are 7 different options for each field. All the participants were told that the left side of the answers represents LIKE and the right represents DISLIKE (except for Color), so the 0 means neutral option.

The 2nd one is open, free answer. This option was hold in case a large amount of participants answered similar words or topics. As we will discuss later, it wasn't possible to take benefit from it.

4.3 Experimental development

As each experiment used 8 pictures and each picture recorded 10 useful seconds of brain wave activity while watching it, each experiment provided 80 seconds of data (80 data). Therefore, it was considered than 20 experiments ($20 \times 80 = 1.600$ data) would be enough data.

For commodity and manageability the participants were actually people from the same laboratory. However, just 10 people collaborated with this project but each of them repeated the experiment a 2nd time some weeks later (so finally it was $10 \text{ people} \times 2 \text{ times} = 20$ experiments).

There was a recommendation to use both male and female so the experiments would be more randomly chosen (not just male or just female). To make it even there were 5 males and 5 females. The names of the participants are included in the thanks page of this project.

4.3.1 Procedure of the experiment

For the experiment 2 laptops were used. One of them would have the role to run the *MATLAB* program (the one that the participant would look at). The other one would be recording the signal waves collected from the EEG device by running the program *MinDSensor5*. So in the first laptop the collaborator should be sitting in front of it, while the person in charge of the project should be sitting in front of the other one.

The idea was that the participant only needed to focus on the pictures and questionnaires. The other in charge should be pausing and starting the *MATLAB* program. As it was considered especially useful to get the data separated from each picture, the *MinDSensor5* had to be also controlled by the person in charge. There are some pictures of how it was managed.



A participant is looking at the picture in the middle of the experiment.



The person in charge needs to pay attention at pause and start the 2 programs.

4.3.2 Extracted data cleansing

As the person in charge can see the seconds that the MinDSensor5 program has been recording, he might take notes of noise, offset values and check if there should be an invalid experiment.

The results from that manual quality filter there where many records that had to be modified because of the mismatch of offset values (bad synchronization between the programs). Also, 2 experiments had to be deleted due to the noise; once because of too much blinking and the other one because there was bad contact between the device and the participant. In addition, 2 picture records from 1 experiment had to be deleted due to coughing.

So finally, the data that was good enough to be used was

$$1.600 \text{ seconds} - 80\text{seconds} * 2 \text{ experiments} - 10 \text{ seconds} * 2 \text{ pictures} = \mathbf{1.420 \text{ valid data}}$$

4.4 Analysis and results

In this section we will see the analysis history that had this project. The contents of this chapter have a combination of analysis and results. The point of this section is to be explained the same way it was thought when working on those analysis so it might be easier for the reader to follow and comprehend the reasons and explanations. So now it would probably be a good idea to continue the explanation separated by small different steps.

4.4.1 Noise detection

The first thing that was performed in the analysis program was the PCA method. As our data have a lot of noise due to blinking and other external reasons, it is necessary to cleanse them. The PCA doesn't eliminate the noise directly; this method just gives us a new perspective of the data converting them to uncorrelated data.

When this is done, the new data have new coordinates but the relative distance between them haven't changed. At the same time that those new coordinates are given, the function that endures the PCA also provides the eigenvalues of the PCA matrix (NxN).

That matrix NxN comes from the N variable matrix where the columns N represent the Frequency. Those eigenvalues mean the weight that each Hz has and, therefore, they say how important will be each frequency for the new values.

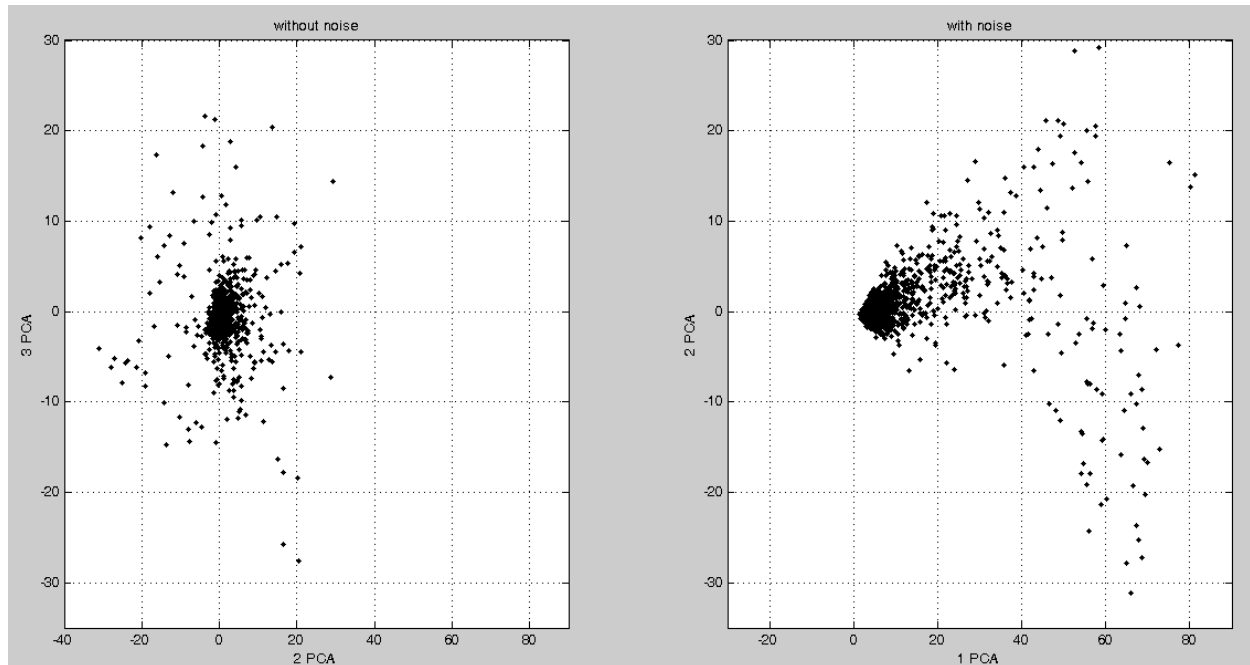
To the right side there is the example that was obtained from the experiment. If we pay attention to those eigenvalues we will see that the first one is wide larger than the other ones. Specifically, it means that the frequency 4Hz would matter 9'5 times ($157'0/16'6 = 9'5$) more than the 5Hz. The 4Hz would accumulate the 75% over the total weight of 19 frequencies.

Thanks to the experience in Mitsukura LAB, it is known that the 4Hz usually come from blinking and it is noise that should not be taken into account. This high value on 4Hz would be very hard to explain otherwise and only noise could get it that high.

lambda <19x1 double>		
	1	2
1	157.0173	
2	16.6218	
3	9.2000	
4	6.6631	
5	4.3929	
6	3.8063	
7	2.3448	
8	1.9821	
9	1.2457	
10	1.1795	
11	1.0792	
12	0.8014	
13	0.6830	
14	0.5677	
15	0.4917	
16	0.4400	
17	0.3548	
18	0.3066	
19	0.2495	

Eigenvalues table result

Once the noise has been detected (thanks to the PCA) it's time to eliminate it from our PCA matrix. That would correspond to the first column of our new 19 column matrix. Let's see some plots and look how different they look.



We can show an example of the all the data collected. The plot with noise is showing the 1PCA against the 2PCA, while the one without noise is just showing the 2PCA and the 3PCA. The one with noise has really wide variance and has a really strange shape on the edges, so it doesn't look like normal data.

4.4.2 File organization

There were a lot of files to deal with. Some of them were the questionnaires answered by the collaborators and some other files were the brain wave measurements. To work with so many variables and files with different conditions wouldn't be easy if there is no previous organization.

So that previous organization would be to transfer all the information from the questionnaires to an Excel. The Excel in “.csv” format can be read by MATLAB and then it would provide wider options to automate the program.

The Excel that was used has the following format:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1			Person				Picture				Results			
2	word 1	word 2	Name	number Person	experiment	order	NUMBER	Windows?	Calm-1, Cool-0	Green?	Green	Color	Windows	OVERALL
3	Artificial		Kanouga	1	1	4	5	1	1	0	2	3	7	5
4	Artificial		Nakamura	4	1	4	4	0	0	1	4	3	1	4
5	beautiful		Takayama	8	1	3	1	1	1	1	2	6	6	6
6	blue		Ogino	5	2	4	6	1	0	0	3	2	5	5
7	cookie	chocolate	Ogino	5	1	2	7	0	1	0	6	7	4	6
8	cool		Kanouga	1	2	1	8	0	0	0	6	1	2	3
9	country house		Kijima	10	1	4	1	1	1	1	6	7	6	7
10	dark	moist	Ogino	5	1	1	8	0	0	0	2	1	3	1
11	depressed		Ogino	5	1	7	4	0	0	1	2	1	3	3
12	feel alone		Nakamura	4	1	3	6	1	0	0	2	1	4	4
13	feel blue		Nakamura	4	1	5	2	1	0	1	3	2	5	4
14	fresh		Ogino	5	1	6	3	0	1	1	6	7	6	7
15	future		Ogino	5	1	8	2	1	0	1	4	3	6	5
16	ghost house		Kiminobu	2	1	5	2	1	0	1	6	1	5	3
17	gray	artificial	Kijima	10	1	6	7	0	1	0	3	3	2	2
18	hard		Ogino	5	2	2	8	0	0	0	2	1	5	4
19	hard		Ogino	5	1	4	6	1	0	0	3	1	6	6
20	interesting		Ogino	5	2	6	2	1	0	1	6	3	7	6
21	like model	like toy	Nakamura	4	1	2	5	1	1	0	2	5	5	5
22	like model		Nakamura	4	1	1	8	0	0	0	2	1	2	4
23	lonely		Ogino	5	1	5	5	1	1	0	3	5	5	5
24	lonely		Ogino	5	2	1	7	0	1	0	5	4	2	5
25	magical		Ogino	5	2	5	4	0	0	1	2	3	3	3

- The first 2 columns correspond to the free answers.
- The next 4 columns contain information about the collaborator; some information to identify the person with an ID number, the experiment (it can be 1st or 2nd) and finally the order in which the picture have been seen.
- The next 4 columns correspond at a similar table that has been seen earlier in this chapter 3 where it identifies which picture it is. An ID number and the description of the 3 different variables.

- The next 4 columns are the results from the questionnaire. Each column represents a row from the questionnaire and as the values in the paper varies from 3 (LIKE) to 3 (DISLIKE) the translation in the Excel is: 7 = 3 (LIKE) and 1 = 3(DISLIKE).

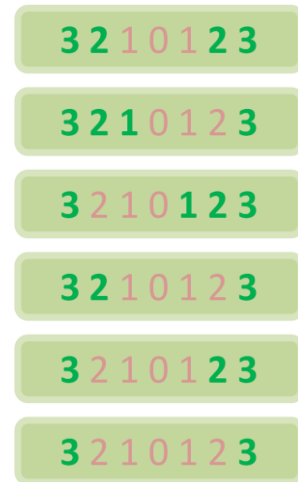
On parallel to this Excel, the files collected from *MinDSensor5* were converted to “.csv” files so they could be read by MATLAB. A small script in the program was in charge to read those files, convert the microvolt data measurements into FFT signal waves and to put all these new transformed data together.

Once the Excel sheet and the MATLAB script were developed, the next step was to enable the possibility of letting the program choose automatically the files. Another script was used to program this new feature. So basically, this new script had an index of all the files with the brain waves organized corresponding to the Excel file. That script would just act as a filter, eliminating all the files or even seconds that might not be desired.

That part of the program was done taking into account that the MATLAB would choose the parameters by itself. That means that those filters had to be variables according to what the program had to run. The automatic variable parts were 2: the subject of the questionnaire that would be taken into account to choose the files, and also the rate following the patterns showed below. (Note: the program it's function is only to check the results after some classification with some few initial files given will be referred in this project as Main Script)

- 6 possible selections (each time the program runs the Main Script only one is valid):

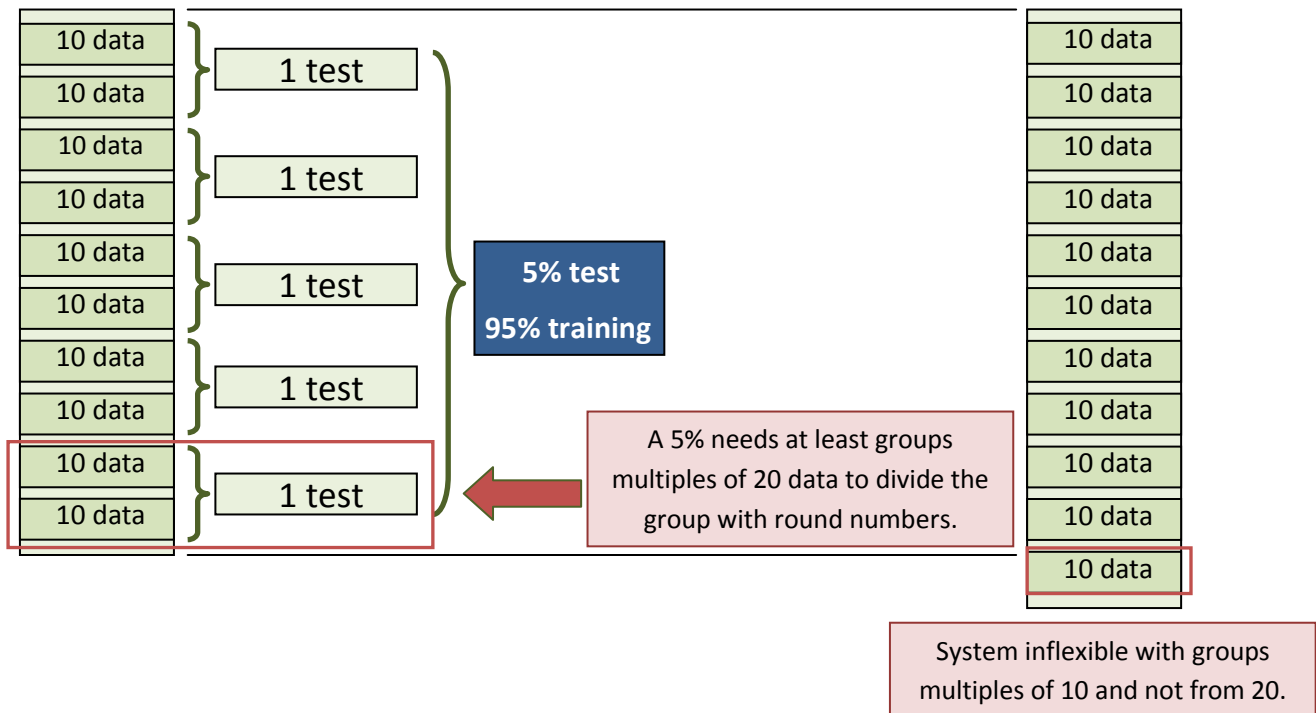
The numbers that are highlighted in green are the grades that actually pass the filter and are taken into account. The files with those red grades are eliminated from the index list.



The idea of the MATLAB program would be to run several times with different starting conditions to discover which ones are the best combinations.

4.4.3 Grouping

This part had to be performed before any classification method could be tested. Usually, the method Cross Validation (leave one out) is used. For this method is common to use 95% of the amount of data as sample data and 5% as testing data. However, in this case it was difficult to carry it out due to the variation of amount of data filtered. The data was organized by groups of 10 seconds for each picture. And those 10 seconds had the same conditions, so the filter chose the files from 10 seconds to 10 seconds. That means that it could be possible to start the main analysis script with 100 data but also it could be that the next one had 110 data. With 100 data is possible to divide by 20 to get a 5% of testing data, but, in the other hand, with 110 data is not possible to get a round number. The solution to this small complication was to use a 10% of testing data in all situations and 90% as training data.



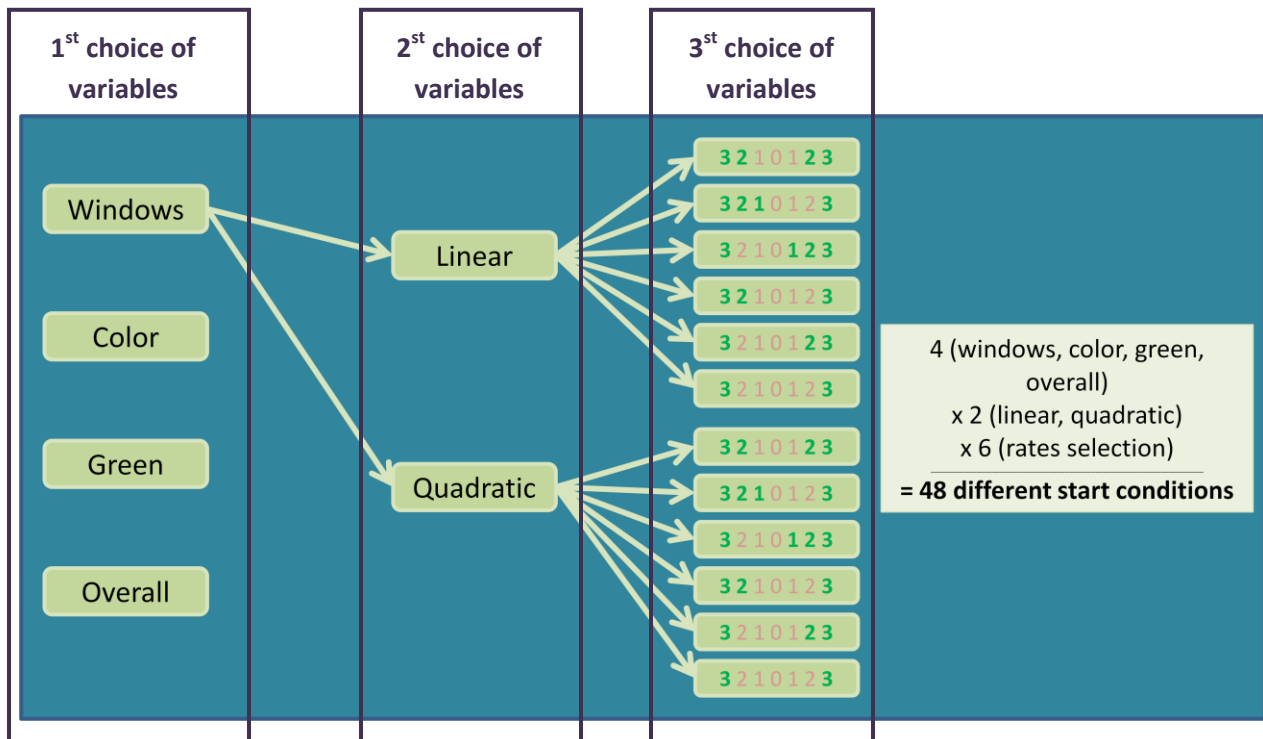
A Cross validation method with multiples of 10 testing data is required in this situation, so this is what is going to be used in the following steps.

4.4.4 Initial conditions - variables set up and filter diagram sequence

The idea was that the MATLAB program could be able to check the results from the classification methods several times per each combination of variables. It means that the main script would run a lot of times with different initial conditions selected by the filter. Although, for each set of initial conditions, the Main Script would also run several times (exactly 100 times) and do an average to calculate the results.

Those 100 times were set up because each time the main program chose different groups due to the Crossing Validation method. The intention was to get an average from all of them and consider the result reliable within those conditions.

To get a clear idea of how many variables and which values or combinations between them the MATLAB program was doing, here there is a diagram that might help to understand it. This diagram is only valid for the LDA classification method, but it will help to understand the idea for all the methods.



- **About the 1st choice of variables.** Basically, there is a lot of data recorded and each data is evaluated by the collaborator with 4 numerical answers. Each answer corresponds to a different topic. In the questionnaire there were those 4 topics (Windows, Color, Green and Overall) and the participant was supposed to answer how likeable that aspect was from his point of view for each picture. So finally, each picture has 4 different topics independents between them. Just one of them can be evaluated at a time (in further studies it could be possible to combine them).

- **About the 2nd choice of variables.** Linear or quadratic. This is specifically from the LDA classification method. When the function is used, the variable of linear or quadratic can be set up. Only one of them can be set up at a time.
- **About the 3rd choice of variables.** In this diagram it's possible to see 2 repeated blocks of 6 rows each block. If we focus only on the first block it's possible to understand that each row represents the answer selected by the participant in the questionnaire. There are 2 types of data, LIKE or DISLIKE. LIKE would be represented on the left side while the DISLIKE on the right side. The green highlighted color means that for that combination the numbers highlighted are chosen. The red colored numbers are discarded in the filter. Only one row can be selected at a time. Those 6 rows and the combinations were discussed and approved by the weekly meetings. It was taken into account that always there should be a difference of 3 numbers in between the LIKE and DISLIKE choices.

So if the whole diagram is read for the first option of each group of variables it would be something like...

- “All the pictures evaluated on Windows, with linear LDA option, that have a punctuation of 3 or 2 on the left side of 0 would be accepted as LIKE data while the 3 or 2 on the right side of 0 would be accepted as DISLIKE data.”

4.4.5 First classification method

As it was the first time to actually get some results, the variables to look at the output were just some basic: the error coming from training data and recognition rate from testing data. The LDA function (*classify* in MATLAB) gives the error produced by the classification of the sample data

directly, so this one was the first output to check. In addition, they gave the recognition results obtained from the testing data. With those recognitions results there is just need to review the labels from each testing original data and then it's possible to get the recognition rate.

As it has been seen previously in the last step, the diagram for the LDA is the following:

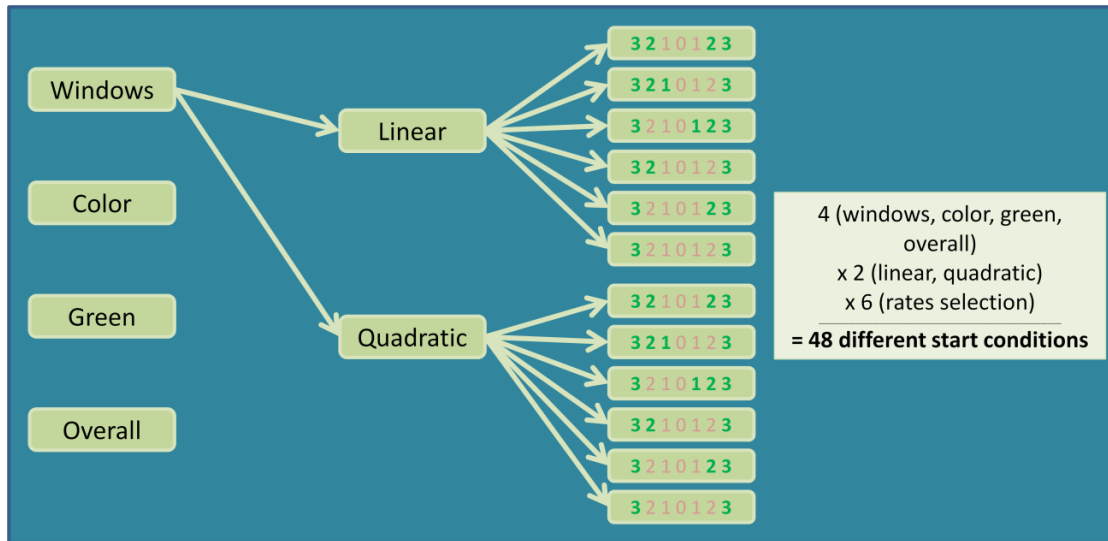
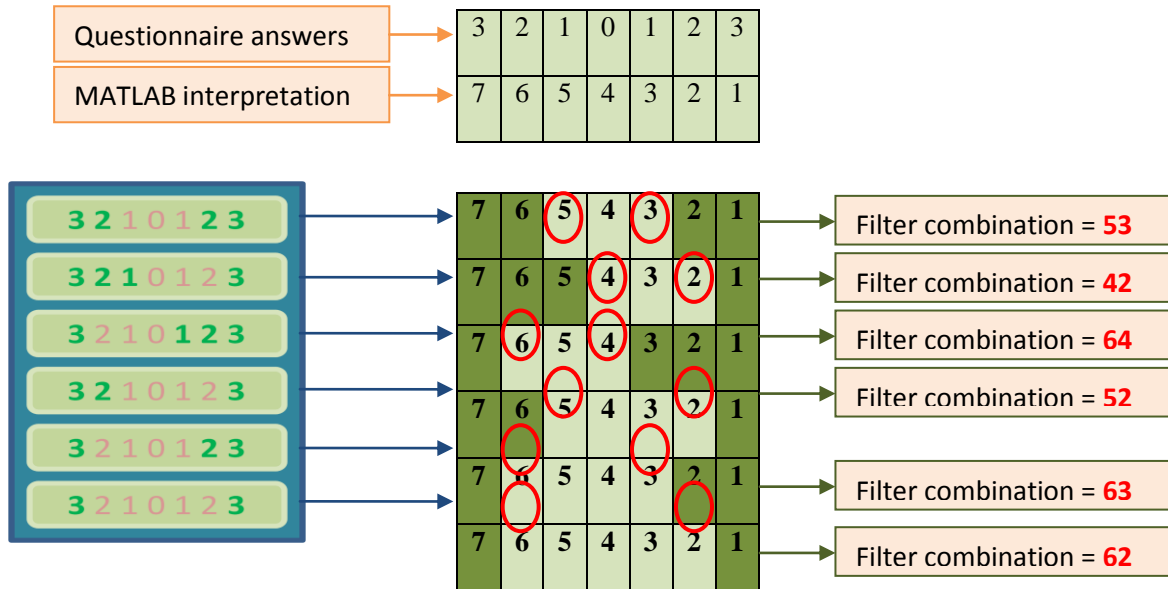


Diagram for the LDA

So there were 48 different starting conditions and one different result for each of them. The results were saved in an Excel sheet where the first 3 columns were the variables selection. The column "filter" is referred to the 3rd group of variables and each number means the following...



The combination shows the numbers that represent the boundaries of the elimination group of data in the filter script.

The other 4 columns are for the results. The first pair is referred to the error that has been committed (with noise and without noise). The second pair is for the recognition rates (with noise and without noise). Each time the Main Script runs, it calculates the classification results for both paths (eliminating the noise and without eliminating it).

So after running the MATLAB program the results were collected in an Excel sheet and arranged from highest recognition rate without noise to the lowest one (the first 24 results are shown).

	A	B	C	D	E	F	G
1	topic rate?	order?	filter	err with noise	err without noise	recogn with noise	recogn without noise
2	Overall	quadratic	42	37.0%	36.8%	80.2%	74.7%
3	Overall	linear	42	40.3%	40.0%	76.0%	73.3%
4	Windows	quadratic	52	40.8%	40.6%	68.1%	67.7%
5	Windows	quadratic	42	43.0%	42.5%	63.5%	63.4%
6	Windows	quadratic	63	38.5%	38.9%	63.8%	63.3%
7	Windows	quadratic	62	38.6%	39.5%	63.4%	62.4%
8	Overall	quadratic	62	39.7%	39.4%	62.6%	62.1%
9	Windows	quadratic	53	41.5%	42.2%	61.5%	60.6%
10	Color	quadratic	62	38.7%	40.8%	62.5%	58.8%
11	Windows	quadratic	64	40.2%	41.0%	59.4%	58.5%
12	Green	linear	63	40.0%	40.2%	56.2%	58.4%
13	Green	linear	62	37.9%	40.2%	57.7%	57.7%
14	Windows	linear	52	42.4%	42.8%	62.6%	56.7%
15	Color	quadratic	63	42.0%	42.3%	56.3%	56.1%
16	Overall	quadratic	63	40.7%	40.5%	55.7%	55.9%
17	Windows	linear	42	43.3%	42.9%	59.6%	55.5%
18	Windows	linear	63	40.0%	43.3%	61.0%	55.1%
19	Color	linear	62	43.8%	43.3%	56.3%	55.1%
20	Windows	linear	62	41.0%	43.5%	60.1%	55.1%
21	Overall	linear	64	42.5%	43.4%	51.5%	54.6%
22	Green	quadratic	42	43.6%	43.4%	58.4%	54.3%
23	Color	linear	63	44.3%	43.6%	54.1%	54.3%
24	Windows	linear	64	40.4%	44.3%	58.6%	54.2%
25	Green	quadratic	52	46.3%	44.3%	59.4%	53.8%

From these results there are a lot of observations:

- First of all, the recognition rates are quite high taking under consideration that the errors are really high. However, if this first observation isn't completely understood (at least for this moment) it's possible to check the others observations.
- It seems there might be a relation between Overall and the combination 42 of the filter and the high recognition rate.
- The next topics with also a great recognition rate are the ones with Windows and quadratic order. It might be interesting to look at the plots.

- Generally, looking at the pair of columns of the error, the one calculated when the noise isn't eliminated is higher than when it actually is. So it makes sense that the elimination of noise is beneficial to the classification.
- Also, when looking for the recognition rate columns, the one with noise seems to have higher rates than the one without noise. This observation is very important because it can make us believe that the last paragraph isn't 100% truthful. There is something missing in these explanations.

Afterwards these observations, there were some recommendations from the weekly reports to look after separating the recognition rate by the different topics. In this case it should be the LIKE and DISLIKE classifications. So in the next time running the program there should be a difference between the recognition rate of LIKE and DISLIKE with noise and without noise. The error wasn't considered anymore.

The next experiment had new 4 columns. The first pair was for LIKE recognitions rates with and without noise and the second pair were for the DISLIKE recognition rates with and without noise.

The results are shown in the following 2 figures.

	A	B	C	D	E	F	G	
1	topic	rate?	order?	combo	recogn LIKE	recogn DISLIKE	recogn LIKE without noise	recogn DISLIKE without noise
2	3 Windows	quadratic	f 62	91.0%	19.5%	91.8%	28.8%	
3	3 Windows	quadratic	b 42	92.5%	13.3%	91.3%	20.6%	
4	3 Windows	quadratic	e 63	91.2%	21.7%	91.3%	30.2%	
5	3 Windows	quadratic	d 52	91.5%	16.0%	91.3%	26.6%	
6	3 Windows	quadratic	c 64	91.2%	21.4%	90.4%	26.5%	
7	4 Overall	quadratic	f 62	90.2%	16.0%	89.3%	21.0%	
8	3 Windows	quadratic	a 53	90.0%	22.2%	88.8%	24.8%	
9	2 Color	quadratic	f 62	86.7%	27.4%	88.2%	26.7%	
10	2 Color	quadratic	c 64	87.5%	19.2%	88.2%	23.0%	
11	4 Overall	quadratic	b 42	87.0%	15.5%	88.2%	28.1%	
12	4 Overall	quadratic	e 63	88.5%	22.3%	88.1%	21.7%	
13	4 Overall	quadratic	c 64	86.9%	25.0%	88.0%	23.5%	
14	1 Green	quadratic	d 52	90.7%	10.5%	87.9%	11.9%	
15	2 Color	quadratic	e 63	86.8%	20.6%	87.8%	24.2%	

Table results arranged by recognition LIKE rate

The previous figure shows the first 15 results arranged from the highest LIKE recognition rates to the lowest ones (without noise).

	A	B	C	D	E	F	G
1	topic rate?	order?	combo	recogn LIKE	recogn DISLIKE	recogn LIKE without noise	recogn DISLIKE without noise
2	1 Green	quadratic	c 64	8.6%	90.9%	11.7%	91.1%
3	1 Green	quadratic	e 63	10.9%	91.8%	14.2%	89.7%
4	1 Green	quadratic	f 62	11.9%	91.0%	15.8%	89.2%
5	4 Overall	linear	c 64	74.9%	33.0%	48.2%	62.7%
6	3 Windows	linear	b 42	83.0%	27.3%	42.8%	62.1%
7	3 Windows	linear	f 62	84.6%	32.2%	46.0%	62.0%
8	2 Color	quadratic	d 52	10.2%	88.6%	42.5%	61.5%
9	3 Windows	linear	e 63	84.2%	34.2%	43.2%	61.4%
10	3 Windows	linear	c 64	83.3%	33.7%	41.8%	60.3%
11	2 Color	linear	c 64	78.9%	33.7%	45.0%	59.9%
12	3 Windows	linear	d 52	83.1%	30.9%	44.3%	59.8%
13	3 Windows	linear	a 53	82.6%	34.3%	39.5%	59.7%
14	2 Color	linear	d 52	77.8%	34.2%	45.4%	59.5%
15	1 Green	linear	d 52	51.2%	43.7%	47.8%	59.1%

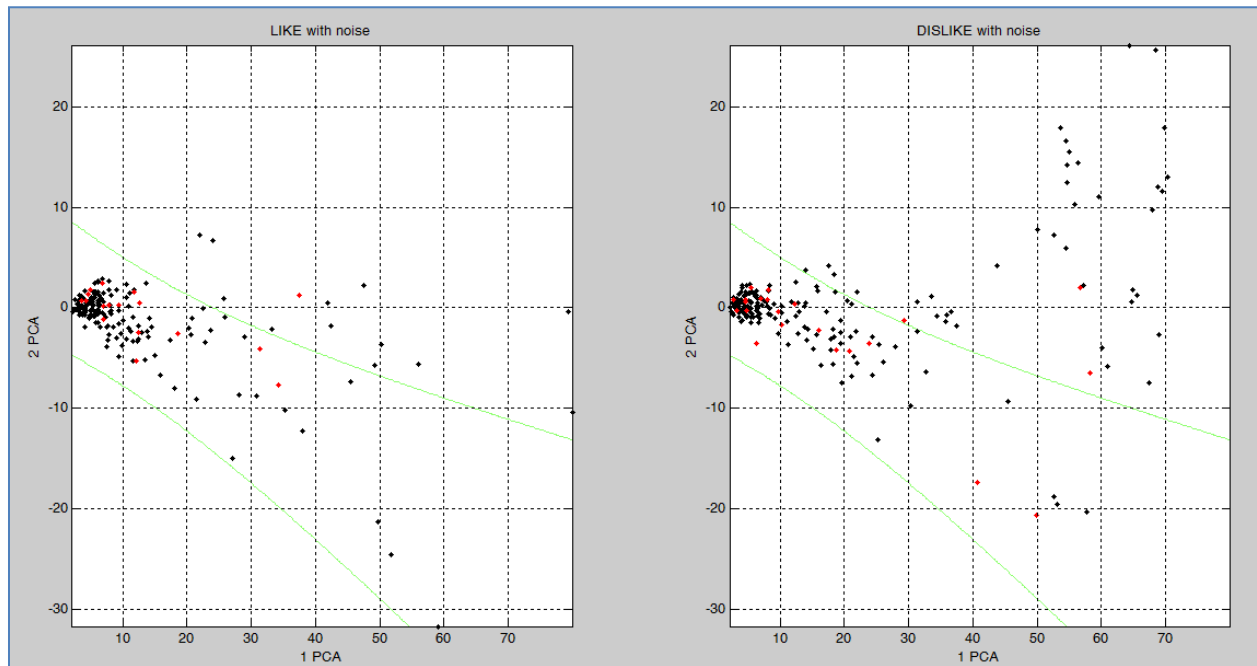
Table results arranged by recognition DISLIKE rate

The previous figure shows the first 15 results arranged from the highest DISLIKE recognition rates to the lowest ones (without noise). From those 2 figures it's possible to get some new observations...

- In the first figure it's possible to see really high values of recognition LIKE rates. They are over 90% and that is a lot. However, if we look at the same time at the recognition rate of DISLIKE it's possible to realize that the recognition rate is almost the complementary of the LIKE rates.
- In the second figures it's possible to observe the same characteristic as in the first figure. The recognition rate on DISLIKE is very high, but the rate for LIKE is almost the complementary to it, so it's kind of strange that this happens with all the rows in both figures. Probably a plot would help out a lot to find out what is going on.
- If the recognitions were reliable, it would be possible to assure there is a relation between the high LIKE recognition rates and the Windows and quadratic variables.

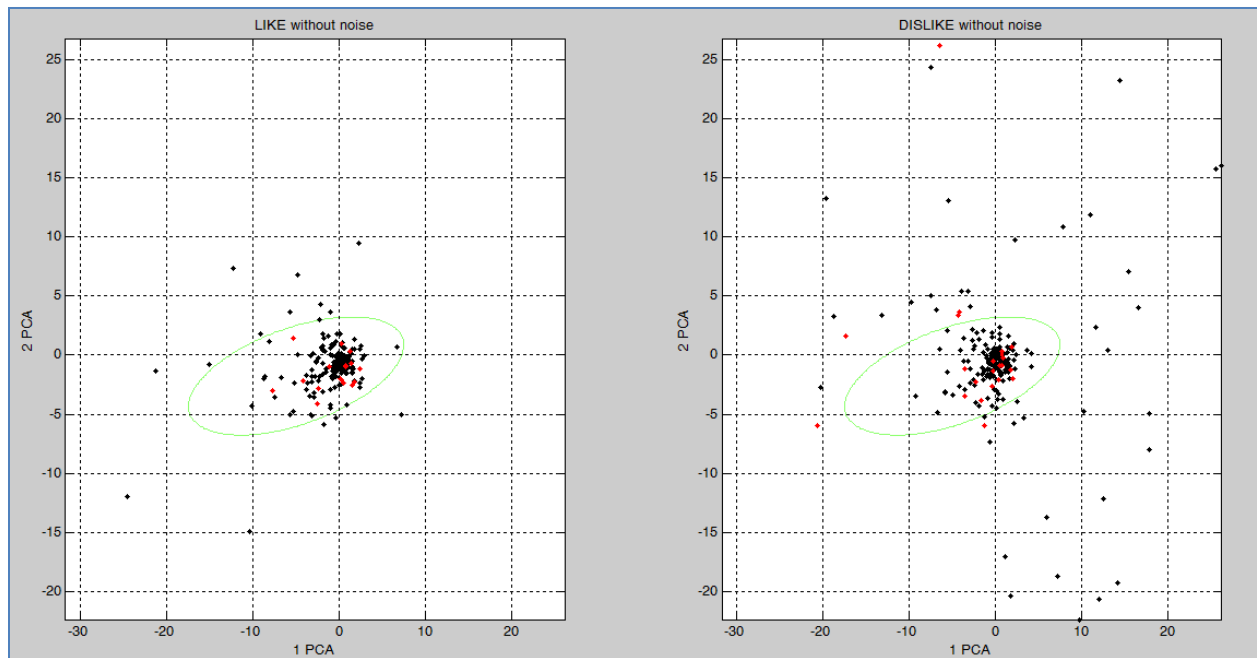
- Again, if the recognitions were reliable, it would be possible to assure there is a relation between the high DISLIKE recognition rates and the Green and quadratic variables.
- There is not much difference between the recognition rates with or without noise in both LIKE and DISLIKE. That makes sense with the 1st figure, where it was impossible to get conclusions about the difference between eradicating the noise or not.

After those Excel tables it would be nice to see a graph. A plot gives a really quick overview of what it is going on. So it was decided, at this point, to plot the 1st and 2nd PCA for each like and dislike columns.



Scatter plot with LIKE and DISLIKE data separated and the LDA function in green color (with noise)

The figure above is from the LDA quadratic method and is showing 2 plots; one from LIKE data (where the noise hasn't been eradicated) and another one from DISLIKE data (where the noise hasn't been eradicated neither).



Scatter plot with LIKE and DISLIKE data separated and the LDA function in green color (without noise)

The figure above shows the same plots as before but without noise. It's possible to see that the LDA quadratic separation line changed its shape between the last 2 figures. The observations from those last 2 figures would be the following ones...

- First of all it has to be noticed that the noise makes the LDA experience some changes on its function. It would also be interesting for helping us to understand what is happening to check the variance from all those plots (it would be done after these observations).
- If we focus on the couple of plots from the last figure it's easy to understand what is happening with the recognition rates and the error. The LDA function is wrapping a vast majority of LIKE data without caring much about the nucleus of DISLIKE data (there are some spread data with high variance, but the majority of DISLIKE data is still in the center). So what happens is that the recognition rates of LIKE data are really high but on detriment of the DISLIKE ones. So there seems to be a correlation between those results: one would be the complementary of the other.

After those observations are done, it was recommended to check the variances first. The Excel sheet of variances results was like this:

	A	B	C	H	I	J	K
1	topic rate?	order?	combo	var 1PCA	var 2PCA	var 1PCA without noise	var 2PCA without noise
2	3 Windows	quadratic	f 62	215.03	23.36	23.36	13.08
3	3 Windows	quadratic	b 42	213.92	19.05	19.05	12.31
4	3 Windows	quadratic	e 63	238.67	24.49	24.49	16.77
5	3 Windows	quadratic	d 52	215.86	19.40	19.40	12.72
6	3 Windows	quadratic	c 64	237.37	24.04	24.04	16.28
7	4 Overall	quadratic	f 62	103.34	8.55	8.55	7.71
8	3 Windows	quadratic	a 53	272.02	26.48	26.48	15.86
9	2 Color	quadratic	f 62	163.22	11.88	11.88	11.15
10	2 Color	quadratic	c 64	235.00	19.77	19.77	16.43
11	4 Overall	quadratic	b 42	206.74	20.74	20.74	9.78
12	4 Overall	quadratic	e 63	121.12	10.41	10.41	8.80
13	4 Overall	quadratic	c 64	167.04	14.75	14.75	11.40
14	1 Green	quadratic	d 52	208.91	24.32	24.32	10.01
15	2 Color	quadratic	e 63	202.81	16.60	16.60	13.30
16	4 Overall	quadratic	a 53	186.20	15.02	15.02	10.59
17	2 Color	quadratic	b 42	236.42	24.69	24.69	11.31
18	4 Overall	quadratic	d 52	179.59	13.66	13.66	9.81
19	1 Green	quadratic	a 53	225.66	24.98	24.98	12.42
20	1 Green	quadratic	b 42	250.75	27.22	27.22	11.72
21	2 Color	quadratic	a 53	225.63	20.24	20.24	14.49
22	4 Overall	linear	b 42	206.74	20.74	20.74	9.78
23	1 Green	linear	b 42	250.75	27.22	27.22	11.72
24	1 Green	linear	f 62	167.22	20.30	20.30	8.83
25	1 Green	linear	c 64	230.02	24.95	24.95	13.07

Organized from
higher LIKE
recognition rate
to lower (without
noise)

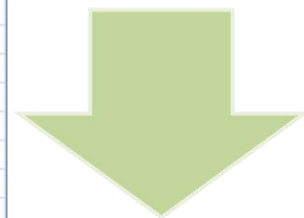


Table results from LDA

We could focus only on the first row that corresponds to the last figures shown that we checked. However, in general, in all the rows it's possible to see how much variance does the data have in the 1PCA if the noise isn't removed. So this result will help the next studies in this project: the noise should be always removed.

All those graphs were done 2 times, once taking into account the first 2 valid PCA (the graphs discussed in here are from this one) and another time with 3 PCA. In any case, the results ended up with the same observations so it would be repetitive to show the graphs and results involving 3 PCA.

4.4.6 Combination of frequencies from the first classification method

The LDA couldn't work very well so it was intended to use another more complex separation method. Although, before appealing to the next classification method, we wanted to give a chance to BE method to check if it the right combination of Frequencies could make any difference.

The BE method was supposed to be applied to the combination of starting variables that provides the highest rates, so we did so. The BE need the starting variables to be fixed. So it was selected the following combination:

- Topic rate? **Windows**
- Order? **Quadratic**
- Filter combination?

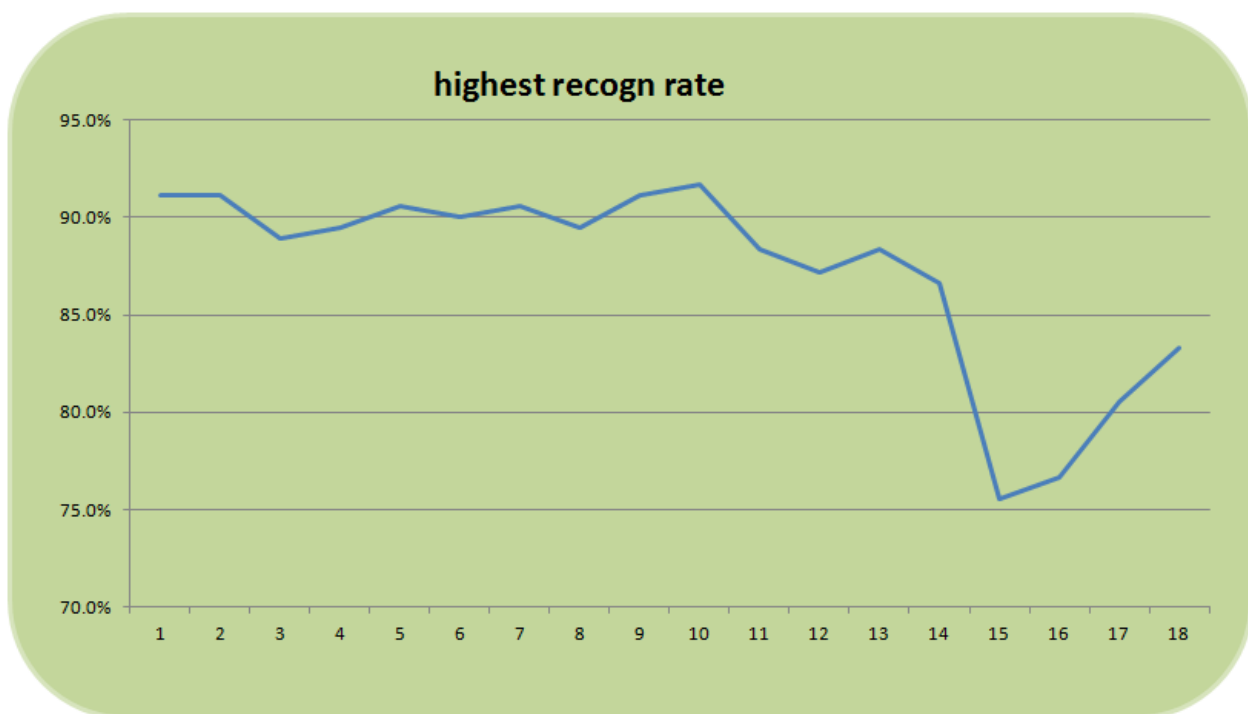
3 2 1 0 1 2 3

Hz	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	like recogn rate
1 step	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	91.1%
2 step	4	5	6	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22		91.1%
3 step	4	6	8	9	10	11	12	13	13	14	15	16	17	18	19	20	21	22		88.9%
4 step	4	6	8	9	10	11	12	12	14	15	16	17	18	19	20	21	22			89.4%
5 step	4	6	8	9	10	12	14	15	16	17	18	19	20	21	22	20	21	22		90.6%
6 step	4	6	8	9	10	12	14	15	16	17	18	19	21	22						90.0%
7 step	4	6	8	9	10	12	14	16	17	18	19	21	22							90.6%
8 step	4	6	8	9	10	12	14	16	17	18	19	22								89.4%
9 step	4	6	8	9	10	12	14	16	18	19	22									91.1%
10 step	4	6	8	9	10	12	14	18	19	22										91.7%
11 step	4	6	8	10	12	14	18	19	22											88.3%
12 step	4	6	8	10	12	14	18	19												87.2%
13 step	4	6	8	12	14	18	19													88.3%
14 step	4	6	8	14	18	19														86.7%
15 step	4	6	14	18	19															75.6%
16 step	4	6	14	19																76.7%
17 step	6	14	19																	80.6%
18 step	6	19																		83.3%

Graph showing BE steps and results

As we were managing 19 variables (frequencies) the BE method has 18 steps. In each step it chooses the best combination in between the frequencies and only one Hz is removed in each step. In the figure above the frequencies eliminated in each step are highlighted in red. At the right side it is shown the recognition rated obtained with this combination of variables.

Basically we should be able to plot the recognition rates and pick the highest one and then look at which frequencies that are taken into account. The graph has been plotted:



Graph from BE plotted results

In this case we already started with really high values and instead of doing a mountain shaped curve it just goes down. This method just makes sense if the bases are right, otherwise it might confuse a little bit more the course of this study.

So it is needed to go back to set the initial conditions again and try to find a better function in order to get reliable results.

4.4.7 Second classification method

As the LDA couldn't fit a proper classification of our data the SVM method was used. To set the MATLAB program it was necessary to think about the diagram of the groups of variables available. So the new diagram was the one as shown:

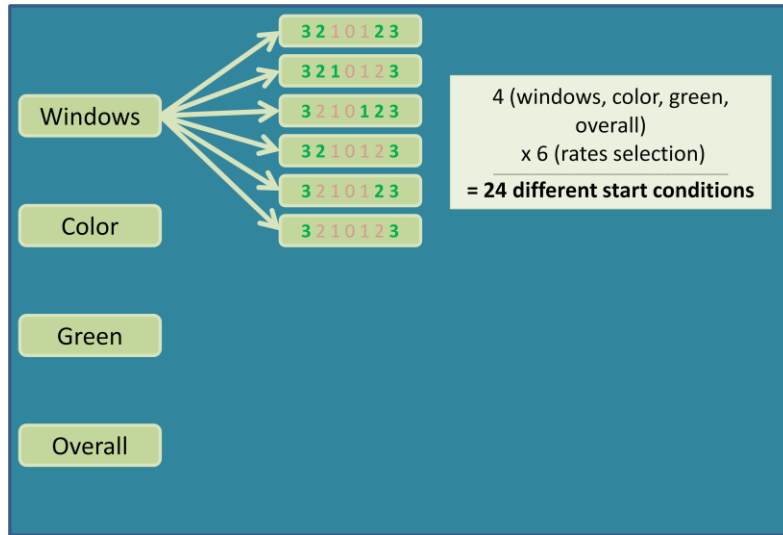


Diagram from SVM

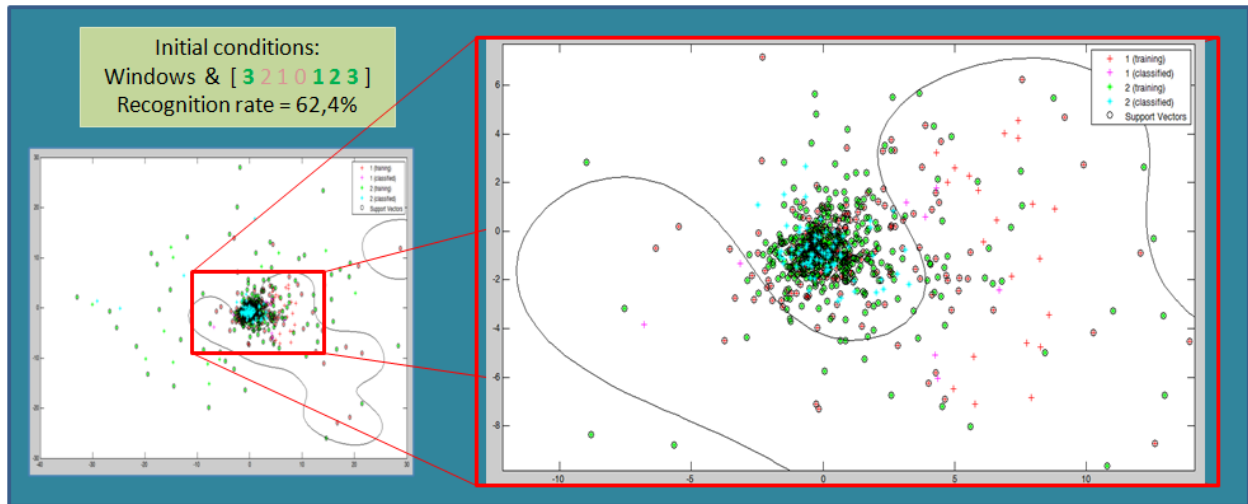
There were 24 different starting conditions. After running this new MATLAB program the new results obtained had 24 rows. Those experiments were done taking into account the first 2 PCA.

	A	B	C
1	topic rate?	combo	recogn rate
2	3 Windows	c 64	62.4%
3	3 Windows	f 62	61.2%
4	3 Windows	e 63	58.8%
5	4 Overall	f 62	58.5%
6	4 Overall	d 52	58.4%
7	1 Green	f 62	55.9%
8	4 Overall	a 53	55.5%
9	1 Green	b 42	54.5%
10	2 Color	d 52	53.9%
11	2 Color	a 53	53.5%
12	3 Windows	d 52	52.8%
13	3 Windows	a 53	52.5%
14	1 Green	e 63	52.2%
15	2 Color	b 42	52.1%
16	4 Overall	e 63	51.7%
17	3 Windows	b 42	51.2%
18	1 Green	d 52	50.9%
19	1 Green	a 53	49.9%
20	2 Color	f 62	48.6%
21	2 Color	c 64	48.2%
22	2 Color	e 63	47.5%
23	1 Green	c 64	47.2%
24	4 Overall	c 64	40.5%
25	4 Overall	b 42	39.2%

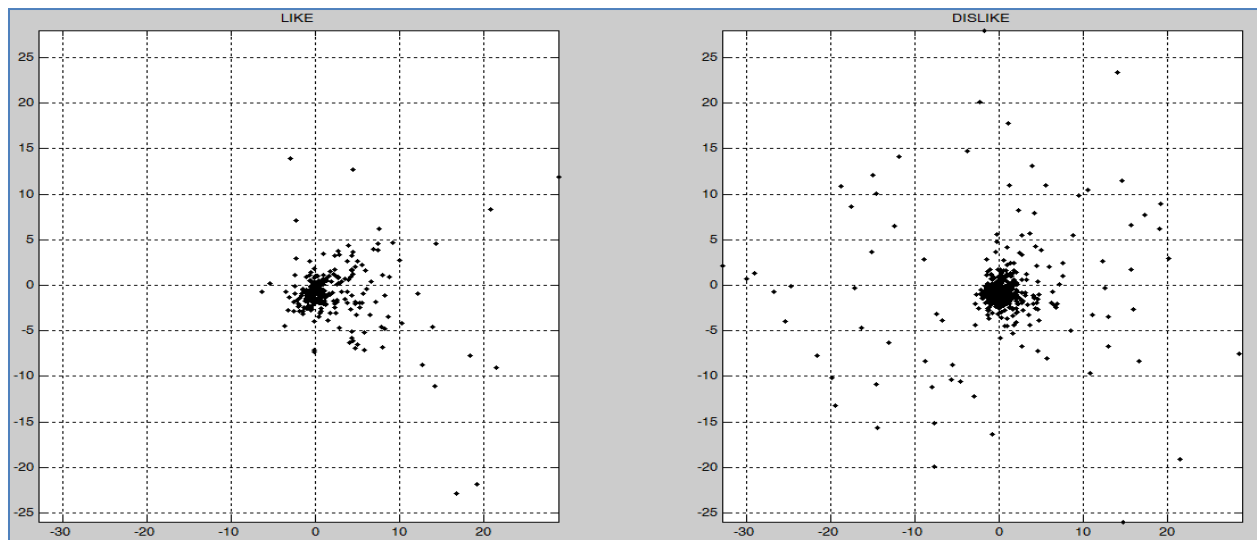
Table results from SVM

On the figure shown on the right it's possible to see those 24 results with their initial conditions and recognition rates. It looks like there might be a relation between using Windows rates and getting high recognition rates. The MATLAB function *SVMclassify* gives directly the recognition rate but it doesn't make difference between LIKE and DISLIKE rates. So it would be a good idea to check the plots of the highest rate and see if those results are reliable or not.

The plots are shown below:



Looking at those plots, if we make a big zoom to the middle where most of the data is concentrated it's possible to see that the SVM function is surrounding the nucleus. However, in this plot is very hard to see which data is from LIKE and which one is from DISLIKE. So it was decided to plot both data (LIKE and DISLIKE) separately.



In the plots above is possible to see the data separated by labels. The axes are in the same scale values and there is a grid to make the reading comprehension easier. Unfortunately, the nucleus

seems to be the same for both data. There are no significant differences on shapes. The boundaries seem to be wider in DISLIKE data, more variance in their locations. Nevertheless, it has to be known that within the initial conditions there are around 250 LIKE data and 670 DISLIKE data taken into consideration. The higher amount of DISLIKE data also may affect on its variance.

So finally it seems there is no way to separate this data. The error in classification is too high and there might be other issues coming from other steps, but not from the classification analysis.

5 Discussion

In this chapter we will discuss the main issues that had been observed in this research.

Summarizing the whole research in few words, we could say that the classifications methods weren't able to find a right combination of variables to separate the data clearly enough so the combination of frequencies found to be important for buildings aren't reliable.

When the first classification method was used (LDA) the results in the plots showed that the data was very close and concentrated. The LDA wasn't able to divide the main nucleus of the data well enough so the recognition rates between both labels were complementary. The one that had the data inside the function had a really high rate while the other one had a really low rate due to counting only the data outside the function. The error was very high, but still there were some hopes relying on other more complexes classify methods.

So the SVM was also tried without success. Even that non-linear method using hyper planes wasn't able to divide the nucleus of concentrated data. The BE method wasn't useful at all without the previous initial variables setting step. So it came out that the main problem was the data.

The experiments weren't realistic enough. The pictures weren't intense enough to make a detectable difference on the signal brain waves. Before doing a future research on the same topic it should be taken into consideration another previous study to find out which kind of pictures (or sight) are good enough to be analyzed.

In any case, the error was the first result that could have explained that from the early begging. So it would be interesting to take into consideration the error as a first step to make sure that the bases are good enough to make future conclusions and applications.

6 Conclusion

Finally, after doing this project we have to recognize that the objectives were not achieved but the path that we followed gave us useful experience to take into account for another possible future research.

- ✓ There was not enough difference between the LIKE and DISLIKE data. It wasn't possible to classify the data correctly even though at the first glance it seems to have high recognition rates.
- ✓ It was not possible to analyze properly the visual aspects that might affect people by reading their brain waves because the bases were not reliable.
- ✓ The error was always too high and the error is basic. The error indicates how reliable the basis is because it kind of means the same as the "recognition rate" among the training data. So if it is too high, there is no way the rates could be high afterwards without carrying a mistake within them.
- ✓ For future research studies on the same direction as this one should take into account that the experiment has to be realistic enough to make a difference in a participant brain signal. If there is not an important impact on the participant, it will never exist any difference between like or dislike data.

7 Bibliography

- [1] Saeid Sanei and Jonathon Chambers. *EEG Signal Processing*. August 2009.
- [2] Jason Wetson. Support Vector Machine (and Statistical Learning Theory) Tutorial. 4 Independence Way, Princeton, USA.
- [3] S. Balakrishnama and A. Ganapathiraju. *Linear Discriminant Analysis – A brief tutorial*. Institute for Signal and Information Processing Department of Electrical and Computer Engineering. Mississippi State University.
- [4] John A. Putman M.A., M.S. *Signal Processing Techniques*. 2007.

Support Internet websites:

<http://www.wikipedia.org/>