Chapter 3

The Spectrum database

3.1 Introduction

In the previous chapter the meaning of depression and some important concepts of the illness, such as how it is assessed, were discussed. This chapter is focused on how the data used in this study was collected, processed and finally used, in this order. These include discussion on the computer vision techniques used to extract features from the subjects' video sequences, which prime example is the use of a model-based approach to the interpretation of the images: the Active Appearance Models.

3.2 Data collection

The aim of this work is analyzing and understanding the behavior of depressed patients. Thus, it is essential to have a set of data to analyze. For building the set of data used in this work, a group of depressed patients were recorded while taking the Hamilton test discussed in Chapter 2.

The process of data acquisition took around two years of work. During this period, the patients were either taking medication or visiting a therapist and they were thus supposed to get better along time. In order to track how their depression severity evolved, they were taking the Hamilton test during an interview with a clinician once a week. Not every interview was video recorded, but just in four different times during their treatment: weeks number 1, 7, 13 and 21.

In addition to the video sequences, a set of personal information is collected from each of the subjects. These are: age, gender, race, marital status, ethnicity and race.

This information is not going to be used directly for classification, however, it could be very useful for a better understanding of the statistical population we are dealing with.

At the end of the data acquisition process, every subject should have been recorded in four sessions of interview. However, some of them quitted the treatment before all the recordings were made and some of them just did not assist to the interview in one specific week. This fact has not implied not taking into account those subjects with an uncomplete number of sessions. They are thus included in the analysis.

A detailed list of the numbers of subjects and sessions that were finally used for analysis could be found in Section 3.4

3.3 Data processing: facial feature tracking

3.3.1 Overview

Once the recordings for all four sessions of each subject are done, we are ready to use them for classification purposes. With no doubt, we cannot use all the information contained in every frame of every session. Hence, it is necessary to approach the problem by some of the state-of-the-art computer vision techniques. These techniques enable us to focus on the information which is useful for us in each of the frames. Since we are aiming to track the behavior of the depressed patients in the video, in this case the useful information is contained on their faces. However, not only a face tracking approach is made, but also a facial feature tracking method. In particular, Active Appearance Models(AAMs) are employed.

In this section, the Active Appearance Models are going to be presented. In order to make the explanation easier to understand, first the Principal Component Analysis method is discussed, as it is a key part of the AAMs.

3.3.2 Principal Component Analysis

Principal component analysis (PCA) is a mainstay of modern data analysis. It is a standard method - used in diverse fields from neuroscience to computer graphics - because it is a simple, non-parametric method for extracting relevant information from confusing data sets.

PCA is a cornerstone method for the linear model construction step of the AAM

[11, 9]. In essence, it is a statistical tool that enables to have a more compact representation of high dimensional data. There are mainly two linear algebra solutions for the PCA: using eigenvector decomposition and using Singular Value Decomposition (SVD). Both techniques rely on the set of theorems detailed below:

1. *A matrix is symmetric if and only if it is orthogonally diagonalizable.*

2. *A symmetric matrix is diagonalized by a matrix of its orthonormal eigenvectors.* Let $\mathbf{A}$ be a square $n \times n$ symmetric matrix with associated eigenvectors $\{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}\}$. Let $\mathbf{E} = [\mathbf{e_1}\ \mathbf{e_2}\ ...\ \mathbf{e_n}]$ where the i^th column of E is the eigenvector $\mathbf{e_i}$. This theorem asserts that there exists a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{EDE}^\top$.

3.3.2.1 Solving PCA using eigenvector decomposition

This first algebraic solution to PCA is based on an important property of eigenvector decomposition. Assuming the data set is $\mathbf{X}$, an $m \times n$ matrix, where m is the number of measurement types and n is the number of samples, the goal of this approach could be summarized as follows:

Find some orthonormal matrix $\mathbf{P}$ *in* $\mathbf{Y} = \mathbf{PX}$ *such that* $\mathbf{C}_Y \equiv \frac{1}{n}\mathbf{YY}^\top$ *is a diagonal matrix. The rows of* $\mathbf{P}$ *are the principal components of* $\mathbf{X}$.

We begin by rewriting $\mathbf{C}_Y$ in terms of the unknown variable $\mathbf{X}$:

$$\begin{aligned} \mathbf{C}_Y &= \frac{1}{n}\mathbf{YY}^\top = \frac{1}{n}(\mathbf{PX})(\mathbf{PX})^\top = \frac{1}{n}\mathbf{PXX}^\top\mathbf{P}^\top \\ &= \mathbf{P}(\frac{1}{n}\mathbf{XX}^\top)\mathbf{P}^\top = \mathbf{PC}_X\mathbf{P}^\top \end{aligned} \tag{3.1}$$

where $\mathbf{C}_X$ is the covariance matrix of $\mathbf{X}$.

According to the second theorem of the previous section our matrix $\mathbf{C}_X$ could be written as $\mathbf{C}_X = \mathbf{EDE}^\top$, where $\mathbf{E}$ is a matrix which columns are the the eigenvectors of $\mathbf{C}_X$. If we choose the the rows of the matrix $\mathbf{P}$ to be the eigenvectors of $\frac{1}{n}\mathbf{XX}^\top$ (in other words, the principal components of the matrix $\mathbf{X}$), we can infer that $\mathbf{P} \equiv \mathbf{E}^\top$. Finally, by adding the well-known theorem *"the inverse of an orthogonal matrix is its transpose"* to the discussion we can finish evaluating $\mathbf{C}_Y$:

$$\begin{aligned} \mathbf{C}_Y &= \mathbf{P}\mathbf{C}_X\mathbf{P}^\top = \mathbf{P}(\mathbf{E}\mathbf{D}\mathbf{E}^\top)\mathbf{P}^\top = \mathbf{P}(\mathbf{P}^\top\mathbf{D}\mathbf{P})\mathbf{P}^\top \\ &= (\mathbf{P}\mathbf{P}^\top)\mathbf{D}(\mathbf{P}\mathbf{P}^\top) = (\mathbf{P}\mathbf{P}^{-1})\mathbf{D}(\mathbf{P}\mathbf{P}^{-1}) = D \end{aligned} \tag{3.2}$$

As we can see, the choice of the matrix $\mathbf{P}$ diagonalizes the covariance matrix $\mathbf{C}_Y$, which was the goal of the PCA. In closing, we can summarize the results of PCA in two bullet points:

- The principal components of $\mathbf{X}$ are the eigenvectors of $\mathbf{C}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$.
- The i^th diagonal value of $\mathbf{C}_Y$ is the variance of $\mathbf{X}$ along $\mathbf{p}_i$.

3.3.2.2 Solving PCA using Singular Value Decomposition (SVD)

This approach of solving a PCA using SVD is a more general method and it leads us to a better understanding of the *change of basis* concept. To begin with, the definition of the SVD decomposition of a matrix will be presented and then the relations between SVD and PCA will be discussed.

Given an $m \times n$ matrix $\mathbf{X}$ whose entries come from the field K, which is either the field of real numbers or the field of complex numbers. Then there exists a factorization of the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^* \tag{3.3}$$

where:

- $\mathbf{U}$ is an $m \times m$ unitary matrix over $\mathbf{K}$, which columns form a set of orthonormal basis vector directions for $\mathbf{X}$.
- $\mathbf{D}$ is $m \times n$ diagonal matrix with nonnegative real numbers on the diagonal, which are the singular values of $\mathbf{X}$.
- $\mathbf{V}^*$ denotes the conjugate transpose of V, an $n \times n$ unitary matrix over $\mathbf{K}$, which columns form a set of orthonormal basis vector directions for $\mathbf{X}$.

If we consider the covariance of the matrix $\mathbf{X}$, a deterministic matrix, to be:

$$\begin{aligned} \mathrm{cov}(\mathbf{X}) &= \mathrm{E}(\mathbf{X}\mathbf{X}^\top) - \mu\mu^\top \\ \mu &= \mathrm{E}(\mathbf{X}) \end{aligned} \tag{3.4}$$

If we assume, without loss of generality, zero mean data and substitute the matrix $\mathbf{X}$ by its SVD decomposition, we get:

$$\begin{aligned} \text{cov}(\mathbf{X}) &= \mathbf{X}\mathbf{X}^\top = (\mathbf{U}\mathbf{D}\mathbf{V}^\top)(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}^\mathbf{t}\mathbf{o}\mathbf{p}\mathbf{U}^\top = \mathbf{U}\mathbf{D}\mathbf{D}^\top \\ &= \mathbf{U}\mathbf{D}^\mathbf{2}\mathbf{U}^\top \end{aligned} \tag{3.5}$$

The goal of the PCA is finding a certain transformation matrix that applied to the data matrix projects it to a subspace where the elements of the data (the columns of $\mathbf{X}$) are completely uncorrelated. In other words, the aim is to obtain a matrix $\mathbf{Y}$ as the result of the transformation of $\mathbf{X}$, which has a diagonalized covariance matrix.

Given the equation 3.5, by multiplying it by $\mathbf{U}$ in both sides and repeating the same process with $\mathbf{U}^\top$, we obtain:

$$\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top = \mathbf{U}^\top\mathbf{U}\mathbf{D}^\top\mathbf{U}^\top = \mathbf{D}^\mathbf{2}\mathbf{U}^\top \tag{3.6}$$

$$\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} = \mathbf{D}^\mathbf{2}\mathbf{U}^\top\mathbf{U} = \mathbf{D}^\mathbf{2} \tag{3.7}$$

Rearranging the left side of the equation 3.7, we finally conclude:

$$\begin{aligned} (\mathbf{U}^\top\mathbf{X})(\mathbf{U}^\top\mathbf{X})^\top &= \mathbf{D}^\mathbf{2} \\ \Rightarrow \text{cov}(\mathbf{Y}) &= \mathbf{D}^\mathbf{2} \end{aligned} \tag{3.8}$$

With this last expression we proof that the matrix $\mathbf{U}$ in the equation $\mathbf{Y} = \mathbf{U}^\top\mathbf{X}$ transforms the data matrix $\mathbf{X}$ in such a way that $\text{cov}(\mathbf{Y})$ is diagonal, which was exactly the goal of the PCA.

3.3.2.3 Practical note

In practice, PCA enables to find a lower dimensional representation of the data which minimizes the squared reconstruction error. Given a set of n samples $\boldsymbol{x}_i \in \Re^d$ (assume, without loss of generality, zero mean data) arranged as a matrix $\mathbf{X} \in \Re^{d\times n}$, PCA finds a matrix $\mathbf{B} \in \Re^{d\times d'}$ such that $\left\|\mathbf{X} - \mathbf{B}\mathbf{B}^\top\mathbf{X}\right\|_2$ is minimum, subject to $\mathbf{B}$ being orthonormal and diagonalizing the covariance matrix $cov\left(\mathbf{B}^\top\mathbf{X}\right)$, with $d' < d$.

The matrix $\mathbf{C} = \mathbf{B}^\top\mathbf{X}$ with $\mathbf{C} \in \Re^{d'\times n}$ is then a lower dimensional representation of the original data matrix. The underlying variables of this new representation are

uncorrelated; if the distribution in the original space is Gaussian, then the variables are independent. The columns of $\mathbf{B}$ are the directions of maximum variance in the original space (typically ordered). These directions are sometimes called modes of variation; they describe a multivariate relation between the variables that decomposes the variability found in the data into uncorrelated components (independent, if gaussianity is assumed—Independent Component Analysis uses higher order statistics to approximate independence with looser assumptions about the distribution of the data).

When implemented, PCA is usually performed using an eigenvalue decomposition of the covariance of the data matrix and keeping the eigenvectors associated with the d' largest eigenvalues, or using the singular value decomposition (SVD):

$$\begin{aligned} \mathbf{X} &= \mathbf{U}_{d\times d}\mathbf{D}_{d\times n}\mathbf{V}^{\top}_{n\times n} \\ \mathbf{B} &= \mathbf{U}_{1,\ldots,d\times 1,\ldots,d'} \end{aligned} \tag{3.9}$$

The value of the eigenvalues kept (and discarded) is directly related to the amount of information kept (and discarded), usually called the energy of the associated data. The total energy is described by the cumulative sum of all eigenvalues. A percentage of energy is said to be kept, by keeping those eigenvectors associated with eigenvalues that make up that percentage of the total energy. When using the SVD, the singular values of $\mathbf{X}$ are the square root of the eigenvalues of $cov\,(\mathbf{X})$, loosely:

$$cov\,(\mathbf{X}) = \mathbf{X}\mathbf{X}^{\top} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}\mathbf{V}\mathbf{D}^{\top}\mathbf{U}^{\top} = \mathbf{U}\mathbf{D}\mathbf{D}^{\top}\mathbf{U}^{\top} \tag{3.10}$$

3.3.3 Facial feature tracking using Active Appearance Models (AAMs)

3.3.3.1 Overview

Active Appearance Models (AAMs) are a model-based approach for the interpretation of images of variable objects. As described by Cootes *et al* [10], an Active Appearance Model contains a statistical model of the shape and grey-level appearance of the object of interest, which can generalize to almost any valid example. However, the Active Model methods like the AAMs are particularly popular in the

field of facial tracking and that is the reason why the discussion along this section focuses on it.

The Active Appearance Model method can be summarized in two phases, which are detailed in the next two sections and summarized below:

- **Training phase:** In this phase the relationship between model parameter displacements and the residual errors induced between a training image and a synthesized model example is learned. Eventually, a model which describes the object to be tracked is built.

- **Fitting phase:** Once the model is obtained, to match to an image the current residuals are measured and the model is used to predict changes to the current parameters, leading to a better fit. A good overall match is satisfied in a few iterations.

3.3.3.2 Appearance model

The models are generated by combining a model of shape variation with a model of the appearance variations in a shape-normalized frame. It is required to have a training set of labeled images, where key landmark points are marked on the objects of interest. In the case we are dealing with these landmarks points should be located on the areas that better describe the shape of the face, such as the eyes, eyebrows, nose, mouth and the outline. In particular, the 66-landmarks shape model showed in Figure 3.1 is used.

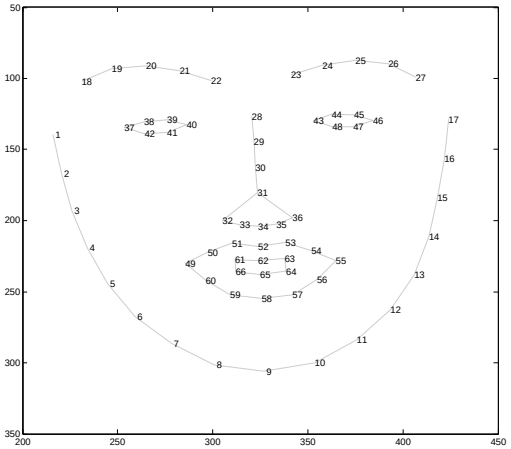

Figure 3.1: Landmark points used for the AAM

3.3.3.3 Model construction

The Active Appearance Model is built from a set of training examples (ie. images of faces) where the position of the landmark points have been marked (typically manually). The shapes are then aligned with respect to each other with an iterative alignment procedure as described in Section 4.4.3.1 to remove rigid motion components such as translations, rotations and scale changes [10].

The x-y coordinates of the points are then arranged in a data matrix. Note that each training sample here is a particular configuration of the x-y coordinates of all the landmark points in the AAM point cloud. For our AAM model, this means the dimension of each sample is $d = 2 \times 66$, that is, 2D coordinates for 66 landmark points. PCA is usually applied, retaining some percentage of the energy. Too many eigenvectors will result in a model that captures noise in the landmark positioning. This can ultimately result in too loose a fit, undermining the robustness of the method. Too few eigenvectors might not capture the whole range of variation of the object.

The mean shape of the point cloud is usually determined. The textured triangles of each of the training images are then warped to the mean shape. This yields a series of warped images in a common reference frame. The pixel data of these is arranged as column vectors, where each row corresponds to a pixel value (which, due to the warp, is supposed to correspond to a specific position on the object). The column space of the data matrix of these vectors is then supposed to describe the full range of possible texture variations for the object. PCA is applied here, resulting in a linear appearance model.

Therefore, any face will be represented as the mean shape plus a linear combination of m shape bases (or basic shapes):

$$\mathbf{O} = \mathbf{S}_0 + \sum_{i=1}^{m} l_i \mathbf{S}_i \tag{3.11}$$

where $\mathbf{S}_i \in \Re^{2\times 66}$ will be the i-th shape basis matrix, each row being the x or y variation of each of the n points (along the columns) for this mode. These matrices are a reshaping of the columns of $\mathbf{B}$, the directions of variance of the model.

The texture (appearance) applied to this shape will be the result of a similar linear combination:

$$\mathbf{A} = \mathbf{A}_0 + \sum_{i=1}^{m'} r_i \mathbf{A}_i \tag{3.12}$$

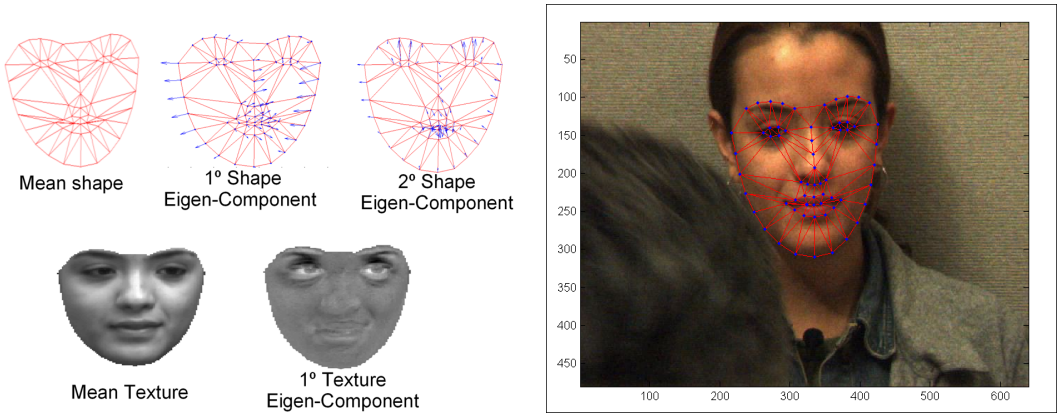

Figure 3.2: Left: AAM model of shape and appearance. Right: Example of an AAM tracking. In blue, the landmarks. In red, the texture mesh.

where $\mathbf{A}_i \in \Re^{1\times p}$ will be the i-th appearance basis matrix, each row being the intensity (gray level) variation of each of the p pixels (along the columns, as discretized for the mean shape) for this mode.

When the PCA transformations are applied, the eigenvalues for each of the eigenvectors are usually kept, as they are in direct relation with the distribution (range) of values of the reconstruction coefficients (ie. the model parameters) found in the training set.

3.3.3.4 Model fitting

To fit the AAM to a new image, an error function of the goodness of fit is minimized by updating the model parameters (conceptually, the PCA reconstruction coefficients of the shape and appearance models). The usual error measure involves the L2 norm of the difference in appearance (pixel value) of the image generated by the model and the original image. Additionally, the parameters $\mathbf{p}$ of a rigid transformation $\boldsymbol{T}_s(\mathbf{p})$ on the shape are found too. (see Section 4.4.2)

$$E(\mathbf{p}; l_{1,\ldots,m}; r_{1,\ldots,m'}) = \left\| \mathbf{A}_0 + \sum_{i=1}^{m'} r_i \mathbf{A}_i - I\left(W\left[\boldsymbol{T}_s(\mathbf{p})\left\{\mathbf{S}_0 + \sum_{i=1}^{m} l_i \mathbf{S}_i\right\}\right]\right)\right\|_2 \tag{3.13}$$

where $I(W[\mathbf{O}])$ denotes the image pixels as warped using the current shape mesh $\mathbf{O}$ to the mean shape in the reference frame used for the appearance basis generation.

The actual fitting process can vary greatly depending on the specific optimization scheme chosen. A full review of all the methods is out of the scope of this document, see [24]. In the interest of speed, it is not a straightforward optimization. Recent methods have explored bypassing usual optimization approaches (calculat-

ing the Jacobian, finding the parameter increment) by regression. Further, the rigid transformation is coupled with the non-rigid deformation, and the whole process is prone to local minima.

3.4 Ready-to-use data: The Spectrum database

3.4.1 Overview

In order to build the Spectrum database, the videos of the interviews of the patients while taking the Hamilton test were processed using the methods explained in Section 3.3. In particular, an AAM-based tracker was applied to all of the videos. As a result of the tracking, the shape and texture information of the face of the patients were extracted.

For this project, only the shape information was used. This information is given by the x-y coordinates of 66 points of the face, in other words, 66 landmark points, as shown in Figure 3.1.

For classification purposes, including the texture (appearance) information would mean with no doubt adding much more and accurate features of the behavior of the depressed patients. Nevertheless, including texture parameters in the experiments would have a big impact in both the computational cost and complexity of the feature extraction process.

3.4.2 Manually labeled data

Although the AAM based tracker has a good robustness in simple scenarios, it is advisable to use another source of reliability. That is the reason why during the tracking of the videos of the depressed patients while taking the Hamilton test, some frames were labeled manually, which means that the landmark points were placed by hand. Note that this data is tracked with the aim of posteriorly analyze it. Particularly, it is going to be used to determine if a patient is depressed or not depressed, which is a difficult task that requires the data to be as accurate as possible. The percentage of data of manually labeled data is around 5%.

Furthermore, this data is used, as shown in Section 5.3, for building a model of the whole set of faces of the database, applying the PCA and trying to reconstruct each of the faces based on the *eigenfaces* obtained after the process.

3.4.3 The missing frames and their implications

As it is easy to imagine, the behavior of the patients is not static enough to make the tracker work in every frame of the videos. The frames where the tracker fails to follow are going to be referred as *missing frames* and the most common situations where they occur are:

- Hand, arm, hair or object occludes the face of the subject.
- The subject's face moves out of the image.
- The subject is continuously doing strange movements with his/her face, such as chewing a gum.

It is very important to find and label where these situations occur along the videos, because the positions of the landmark points during those frames are unreliable and the features cannot be extracted from them.

The bad news about these frames that are missing by the tracker is that they cannot be automatically found. They have been labeled by undergraduate students of the University of Pittsburgh, as part of a collaboration with the Affect Analysis research group of this same institution.

The missing frames will be removed from every video sequence before the analysis. This will result in the occurrence of *jumps* in the sequence, as not all the frames will be consecutive. As explained in Chapter 5, when extracting the features and computing the difference of the position of the landmarks between consecutive frames, this jumps will be taken into account, as they will be avoided.

3.4.4 The Spectrum database in numbers

As discussed in the introduction of this section, the AAM-based tracker is used for extracting the shape and texture information -although the latter is not used in this project- from the videos of the patients. Once this process is over, the reliability of the tracker is insured by a carefully visualization of the videos and the missing frames are also found.

The total number of sessions and subjects present in the clean version of the database are shown in Table 3.1. They are organized by sessions 1, 2, 3 or 4, corresponding to the different weeks in which the patients had the interview with the clinician: 1st, 7th, 13th and 21st week from the beginning of their treatment.

# subjects	Sess. 1	Sess. 2	Sess. 3	Sess. 4	Total # sessions
52	48	39	33	29	149

Table 3.1: Total number of subjects and sessions of the Spectrum database.

Note that the number of sessions is not even because not every subject went to every session of interview, as it has been explained in Section 3.2.

3.4.5 Interviewer variability

When humans converse, we adapt multimodally to one another. Coordination between speakers' and listeners' head movements has been widely reported (Bernieri, Davis, Rosenthal, & Knee, 1994; Cappella, 1981; Condon, 1976; Lafrance, 1985) and [31]. Even the sex of the speaker-listener has an implication on this coordination [23]. Thus, the ideal situation would be having a single interviewer for every patient. Otherwise, it is expectable to have a certain bias of the patient's behavior depending on the person they are speaking with. Unfortunately, this is the situation we have in the Spectrum database, in which the videos were generated with four different interviewers.

As the number of sessions corresponding to each of the interviewers is not enough to analyze them separately, the only way to address this problem would be analyzing the interviewer/patient behavior at the same time. In fact, the interviewer was also video recorded during the sessions, but at the time of this project they have not been tracked and cannot be analyzed.

Despite the implications that this variability could have in the classification process, the inter-subject variability due to the change of interviewer will be ignored. Taking it into account could certainly be a source of interesting results and it is thus part of the future work of this project.