# Chapter 7

# Conclusions and future work

## 7.1 Conclusions

Facial expression analysis has become increasingly possible in the last years, mostly due to the advances in machine learning and computer vision techniques. But nobody said it was easy. Even approaching a problem in which it is not necessary to detect or recognize specific facial expressions, but only extract information regarding the general behavior of the face, becomes a very challenging work.

The main aim of this project was to build a first approximation of a classifier able to determine wether a person is depressed or non depressed. Although the classification results do not exceed the 76%, they appear to be very promising. In order to improve them, every step is important: choosing the data samples to train the classifier – depending on if we want it to be able to classify evident or subtle examples –, the alignment process – enables to compare the features from different subjects –, the feature extraction method, the classifier and the training-testing scheme to determine the goodness of the tests.

Dealing with big amounts of data, such as the video sequences from the Spectrum database (20,000 frames long in average) is always difficult. In this project the adopted solution was segmenting the data in non-overlapping windows, turning out that the 10 seconds long windows gave the best results. However, since this leads to a data reduction, a lot of information is lost in the process.

To the inherent complexity of the tasks discussed above, the difficulty of understanding and finding the good features to extract sums up. In this framework, the stretch collaboration with expert psychologists was essential, since they are able to point out which specific parts of the face are worth to pay attention to.

Perhaps what can be surmised from these tests is that it is not at all clear where

the useful information resides. If a human being is able to determine in most of the cases if a person is depressed just by a simple visual analysis, why not taking for granted that a machine can eventually be trained to perform the same task?

## 7.2 Future work

One fascinating thing about research is that, although the topics are normally very specific, work on them can be endless. Even more when it is about the Spectrum database and when this database is growing. The following list describes the data available or that will be eventually available.

- **Tracked video sequences of the patients:** The AAM tracking of all the subjects of the database is now completed, making a total of 149 full sessions of interview.

- **Tracked video sequences of the interviewers:** The AAM tracking of the interviewers has not yet started, although is one of the priorities for the evolution of the project. In a diadic conversation, the correlation between the subjects' behavior is enormous [23]. Thus, analyzing the movements of the interviewers is essential for a full understanding of the behavior of the patients.

- **Audio recordings for the patients and the interviewers:** Many relevant information can be inferred from the speech analysis. The pauses between words and between each person's turn can give us information about how sever is the depression of the patients. Furthermore, a fundamental frequency analysis can be done.

- **Transcriptions of the interviews:** At this time, students from the University of Pittsburgh are working in making the transcriptions of all the interviews recorded in the database. These transcriptions, in conjunction with the speech information, give rise to many possible studies.

- **Temporal segmentation of the interviews:** In order to use the speech information, it is fundamental to have the temporal segmentation of the conversations. It is part of my present work to achieve an automatic speech segmentation system. Its challenge relies on doing it based on the tracking information (landmark points), instead of directly from the audio signal.

- **Gaze tracking:** Applying state-of-the-art computer vision algorithms to track the gaze of both the patients and the interviewers is part of the future work of this project. With this information we will be able to analyze the eye-contact between them, as well as the amount of time the patient looks towards the interviewer.

Taking for granted that all the described information will be eventually available, the Spectrum database will be probably the most complete database in the depression field. In the following points some of the future work lines are described:

- **Feature selection:** As seen in Chapter 6, the accuracy of the classification is highly dependant on the landmark points of the face chosen for the analysis, which means that some of them are more relevant for the depression assessment than others. Furthermore, statistical features have been used as the input of the classifier, but it has not been determined which of them are the most discriminant. A feature selection method can be applied by running the classification using one statistical feature at a time, leading to a data reduction and also to the investigation of which other features could be used in order to improve the classification results.

- **Extracting features from the speech:** Previous research has shown the relation between vocal timing in face-to-face interaction and depression [31]. Once the speech segmentation of the interviews of the database is finished, it will be possible to extract the amount of time the patients spend before answering a question, or how long are the intervals between words. One of the symptoms of depression is having less reaction capability and these two features can be a good approach to quantify this characteristic.

- **Classifying mild cases of depression:** Among all the sessions of the Spectrum database, the most extreme cases of depression were chosen as data for the classification process, following the criterion described in Section 2.2.3. However, it is important to be able to classify the mild cases. This can be achieved, for instance, by a better understanding of the features that contribute the most to the training step, this is, by successfully applying a feature selection method. Another possibility is to complement the features by including the information extracted from the speech, as explained in the previous point.

- **Analyzing the patient-interviewer interaction:** This analysis can be done once the AAM tracking for the video recordings of the interviewers is completed. Since there is not one single interviewer for all the subjects and the behavior of the patients is believed to be coupled to that of the interviewer, this information might improve the classification performance. For instance, one interviewer can induce a more talkative interview, or with more head motion [23]. However, some efforts have to be made on finding a way to use this information as input features of the SVM.

- **Using head motion angles as features:** The AAM tracker employed in this project is a 2D tracker. However, two cameras, one in each side of the person, were used to record both the patient and the interviewer, with the initial intention of performing a 3D tracking. In practice, both 2D and 3D trackers were applied to every subject of the database. The first steps of this project, not described in this report, were focused on the 3D tracking, and in this context we realized that the depth information ($z$ coordinate) was not as accurate as expected. This would not be a big problem if it was not for the necessity of performing an alignment step before the feature extraction procedure (see Section 4.4). When applying Procrustes, as the three coordinates are taken into account, the *false* depth information caused deformations of the face that in many cases were unacceptable. That is the reason why the 2D tracking was used instead.

  According to Dr. Jeffrey Cohn, head of this project, head motion is presumed to be related to depression severity. Particularly, the pitch angle (up and down movement) is the most relevant among the three parameters (pitch, yaw and roll). If we had the three coordinates for every landmark point of the face, we could compute the head motion in every frame using the rotation matrix resulting of the Procrustes analysis (alignment). However, since we are presently using the 2D tracker, the head motion extracted by this method only shows these angles in the projection into the plane, and we consider them as not being accurate enough for using them as features for classification.

  It is thus part of the future work of this project to investigate a more intelligent way to compute the head motion angles, since the 3D tracking is unlikely to be improved in the short term.

- **Studying the relation between speech and depression:** Once the tran-

scripts of the interviews are completed and the work on the video-based speech segmentation has succeeded, it will be possible to detect the time interval in which a patient says a specific word. In most of the cases, the first session of a patient of the Spectrum database corresponds to a case of severe depression, while the last one corresponds to a non-depressed status. This means that information of the same subject being depressed and non-depressed is available. Since the Hamilton test is normalized, there are many words that are likely to appear in both sessions and it will be possible to compare the way a given subject says a word when he/she is depressed or not.

- **Using Action Units as features:** Members of the Human Sensing Lab of Carnegie Mellon University, where this project was developed, have been working for two years in building an Action Unit recognition software [30]. The Facial Action Coding System (FACS), devised by Ekman and Friesen (1978) [14] is a comprehensive system that measures all visible facial muscle movements, and not just those presumed to be related to emotion.

  Many psychology papers analyze the presence or absence of certain Action Units, as well as other statistical measures related to them, to study specific psychological disorders. Action Units have been even used successfully for suicide reattempt risk assessment [2] and their use in the depression field is not new. As an example, in [20] the response to comedy (a positive stimuli) of depressed and non-depressed subjects was analyzed.

  Manual FACS coding is very costly, since it has to be made by professional coders. However, once the Action Unit recognition software is available, the whole Spectrum database could be coded automatically or just with little supervision. Using the Action Units performed by the subjects as features for classification is believed to give very promising results.

Preliminary results on some of the work described above, as well as the classification results presented in this report have been gathered in a paper entitled *Detecting depression from facial actions and vocal prosody*, accepted in the International Conference on Affective Computing and Intelligent Interaction (ACII 2009) [7].