Chapter 5

Features for classification

5.1 Introduction

If the tracking is sufficiently robust and precise, it is conceivable that using only shape information would be enough to classify the depressed and non depressed class of subjects. It is certain that adding the appearance information will describe more accurately the morphology and expressions of the face, but it is also certain that almost every change in appearance (even wrinkles, shadows, etc.) will result in movement of the landmark points as the AAM model is fitted.

There are several features that make shape-only classification attractive. It is conceptually simpler, and cleanly divides the problem into more manageable parts. The tracking subsystem can benefit from advances in computer vision tracking (a huge field in itself), giving the classification subsystem some independence from the input sequence.

It is possible to use the whole shape of the face as a features, this is, being the position of the landmark points the features for classification. Nevertheless, if we check out the numbers we are dealing with:

- Each face has 66 landmark points.
- Each session has an average of 20,000 frames.
- There are 149 sessions in the Spectrum database.

This makes a total of 196,680,000 two dimensional landmark points, a huge amount of data.

Being the kind of data we are working with very difficult to collect, a study like the one we are developing along this project has never been done. This is an

injection of motivation for the researcher, but at the same time makes the problem very challenging. No clue is given to us. However, analyzing the problem we can conclude that there certainly are a set of landmark points that will not give us any hint about how depressed a person is. This set of points if formed by the outline of the face (jawline) and the nose and are therefore going to be discarded as features for classification.

Even discarding the outline of the face and the nose, which correspond to 17 and 9 points, respectively, we still have a huge amount of data to deal with. It is thus needed to find a more intelligent way to perform a data reduction procedure.

5.2 Data reduction: statistical analysis

As a data reduction procedure, a statistical analysis will be chosen. In other words, instead of using the position of the landmark points themselves, certain statistical measures will be used.

As the statistical measures do not have at first hand an intuitive meaning and as we do not know the variability of which part of the face is more discriminative to assess depression severity, one first approximation for building the features for classification will be taking a combination of them, and organizing them hierarchically.

In particular, three layers of features will be defined. The final goal is to obtain only one feature vector for each session used for classification. These three layers are described in detail in the following sections.

5.2.1 Frame level

The statistics are computed on the amount of variability between consecutive frames. This is, what is of our interest is how much the face changes from one face to another, which can be interpreted as a measure of velocity.

The frame level thus consists in taking the sequence of frames of a given session and computing the difference between the consecutive ones. Note that as explained in Section 3.4.3, after removing the *missing frames*, not all the frames of the resulting sequence are consecutive (there is a time jump between them). These frames will be ignored as they do not provide a realistic measure of the movement of the subject.

The result of the frame level of features is a sequence of vectors, each one containing the *velocity* of the landmark points, i.e. the difference of position of each one of the landmark points between one frame and its previous one.

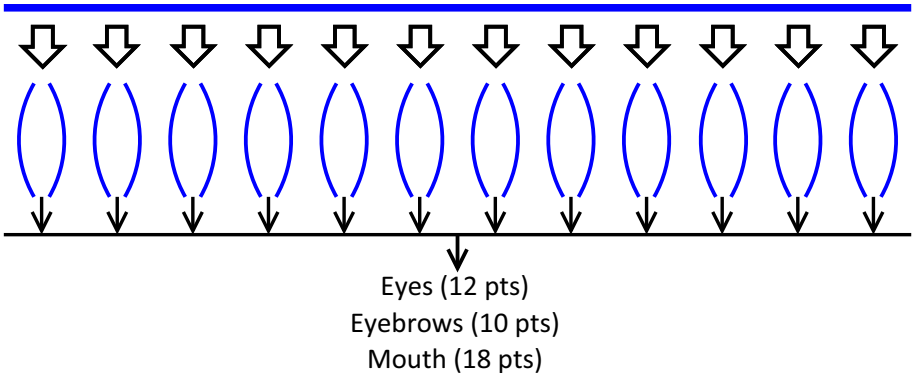


Figure 5.1: Feature extraction: frame level.

5.2.2 Group level

In order to build the group level, two steps are followed.

Firstly, the sequence resulting from the frame level is segmented into non-overlapping windows. The main idea is to be able to capture the movement of the subject in small time intervals. The optimum length of the window is unknown a priori, but it could be determined after analyzing the results of the classification. In particular, the simulations will be run using three different segmentations: 150 frames, 300 frames and 600 frames, which correspond to windows of 5 seconds, 10 seconds and 20 seconds, respectively.

Secondly, the goal is to obtain one feature vector for each of the windows of the segmentation process. For this purpose, a group of statistical functions will be used: the mean, the median and the standard deviation.

If we take into account the definition of the frame level, the statistical measures used in this layer are applied to the velocity of the landmark points. Therefore, for each of the landmark points the mean, median and standard deviation of its velocity are obtained. Then, by concatenating these statistical measures for every landmark point we get a single feature vector for a given window of the sequence. Thus, the result of this level of feature extraction is one feature vector for each of the segments of the sequence.

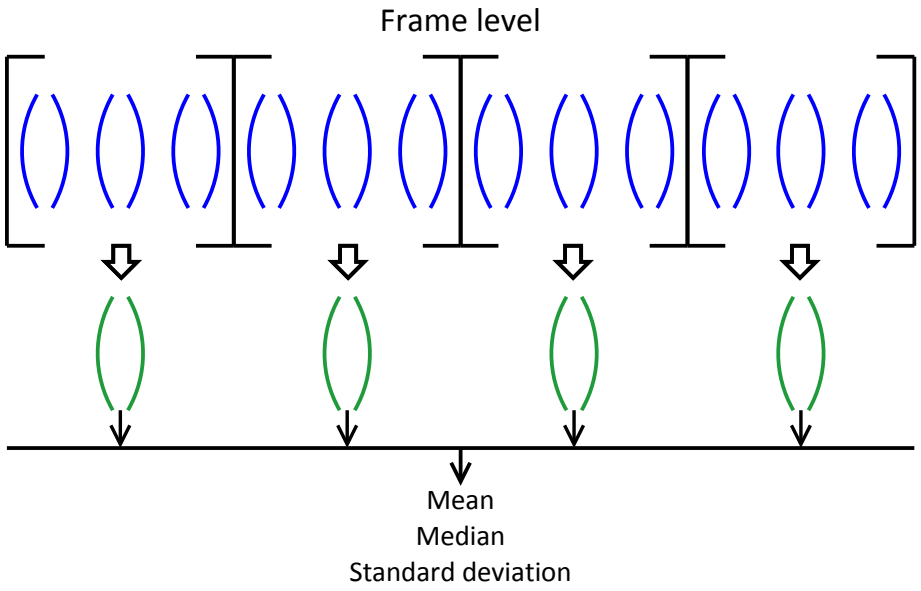


Figure 5.2: Feature extraction: group level

5.2.3 Global level

As a final step in the feature extraction process and keeping in mind that the final goal is to obtain a single feature vector for the whole sequence of study, another set of measures is applied. On one hand, two statistical measures: the median and the mean. And on the other hand, the maximum and the minimum.

The minimum, maximum, median and mean are applied to the statistics resulting from the group level, i.e., there are 12 possible combinations among them. By concatenating all these combinations of measures we obtain one single feature vector for the whole video sequence and the feature extraction process comes to an end.

As it is easy to imagine, an interpretation of, for instance, the mean of the median of the velocity of a landmark point, is difficult to give. Nevertheless, and although this research might be interpreted as being interdisciplinary (computer vision and machine learning on one hand and psychology on the other hand), its main goal is to build a classifier capable of separating the depressed and non depressed class of subjects, but not understanding which are the features that enable to distinguish them. With no doubt, it would be extremely interesting if this project could help to characterize the human behavior in a depressed status, but the nature of the features used makes it a very difficult task.

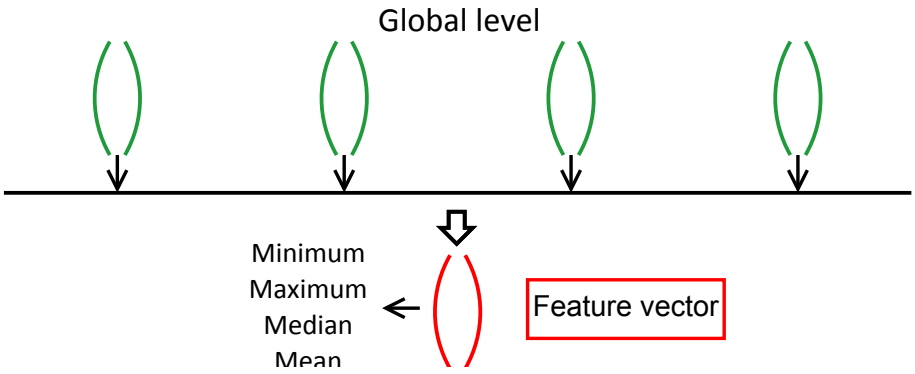


Figure 5.3: Feature extraction: global level

5.3 Using shape coefficients as features

5.3.1 Overview

Along the previous sections a data reduction method was presented. Although the whole explanation was focused on reducing the big amount of data of a complete shape of the face, comprised by the position coordinates of 66 landmark points (40 discarding the nose and the jawline), this approach can be applied to any type of data, for instance, the so called *shape coefficients*.

Loosely speaking, the shape coefficients of a given face shape describe its variation of movement with respect to a certain mean shape. They will give us an idea of the amount of energy a given face has, which it is believed to be a discriminant measure between the depressed and non depressed subjects. In order to extract them, a model of faces has to be built, i.e., a set of vectors defining a certain subspace. The coefficients needed to reconstruct a given face using this model is what we call the shape coefficients. In other words, they are the scalar numbers by which the vectors of the subspace are multiplied to represent a given face.

5.3.2 Building the model using PCA

According to Section 3.3.2, PCA allows us to represent a given set of data, $\mathbf{M}$ in a subspace where its components (columns) are completely uncorrelated. This subspace is defined by a set of vectors, which is what we call the *model* of the data. Each one of the data examples can be expressed as a linear combination of these

vectors and if only the vectors with higher weight are kept, in other words, the ones with higher energetic contribution, we are able to achieve a data reduction.

The model of faces built has to be able to represent as precisely as possible each one of the faces of our database. The ideal situation would be thus to take into account every frame of every session of the database, but as we have earlier discussed, this means dealing with a huge amount of data. The frames that were manually labeled during the tracking process (see Chapter 3) are used instead. As these manually labeled data contains frames for every session of the database, we assume it is a good representation of it.

In our particular scenario, the matrix $\mathbf{M}$ is a $132 \times N$ matrix, where N is the total number of manually labeled faces used for building the model. Each of the rows contains the coordinates of the 66 landmark points of one of the faces, being the x and y coordinates placed one after the other. If we now consider the matrix $\bar{\mathbf{M}}$ as to be a $132 \times N$ matrix which contains in each of the columns the mean face of the manually labeled data, a matrix $\mathbf{m}$ is defined as:

$$\mathbf{m} = \mathbf{M} - \bar{\mathbf{M}} \tag{5.1}$$

$$\mathbf{M} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,66} & y_{1,1} & y_{1,2} & \cdots & y_{1,66} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,66} & y_{2,1} & y_{2,2} & \cdots & y_{2,66} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,66} & y_{N,1} & y_{N,2} & \cdots & y_{N,66} \end{bmatrix} \tag{5.2}$$

In summary, each of the faces contained in $\mathbf{M}$ is subtracted the mean face of the set of data. This mean face plays a very important role during the reconstruction of the data, as it is shown in the next section.

As a final step, a PCA over the matrix $\mathbf{m}$ is solved using Singular Value Decomposition (see Section 3.3.2). In practice, this enables the factorization of the matrix as:

$$\mathbf{m} = \mathbf{U}\mathbf{D}\mathbf{V}^* \tag{5.3}$$

$\mathbf{U}$ is a 132×132 matrix containing the model of our data (set of orthonormal vectors). Note that $\mathbf{U}$ and $\mathbf{V}$ can be interchangeable just by transposing the matrix $\mathbf{m}$ before the SVD decomposition. This model is afterwards used for reconstructing each of the faces of the database. This process is explained in detail in the next section.

5.3.3 Reconstruction of the data

The columns of $\mathbf{V}$ define a subspace which contains each one of the faces used for building the model (manually labeled faces). The representation of one of these faces in this subspace has the advantage of being uncorrelated with the representation of another one. To find this representation the matrix $\mathbf{m}$ is projected into the subspace, which can be done just by multiplying it by $\mathbf{V}$. The result of this projection is a matrix $\mathbf{C}$ containing in each of its rows the *shape coefficients* of the corresponding face.

$$\mathbf{C} = \mathbf{m} \cdot \mathbf{V} \tag{5.4}$$

Each face has associated 132 scalar values, the shape coefficients. These values, multiplied by the model $\mathbf{V}$ lead to a perfect reconstruction of the original face.

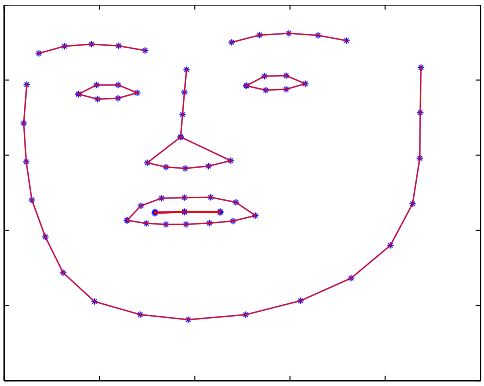
$$\mathbf{m} = \mathbf{C} \cdot \mathbf{V}^{\top} \tag{5.5}$$

In order to complete the reconstruction of the face, the mean face has to be added: $\mathbf{m} + \bar{\mathbf{M}} = \mathbf{M}$.

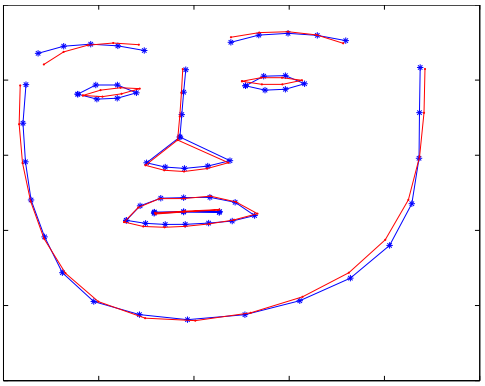
If the face that is being reconstructed by the shape coefficients has been part of the set used for building the model $\mathbf{V}$, the reconstruction is perfect. However, only the manually labeled faces took part of this process and our interest focuses in being able to reconstruct every face of the database. As mentioned in previous paragraphs, as the manually labeled data come from the same subjects/sessions as those faces we want to reconstruct, we assume that using the model $\mathbf{V}$ for their reconstruction would cause a small error that we accept to include in the process. Figure 5.4 shows a reconstruction example.

The reconstruction error can be computed as the matrix norm of the difference between the original face and the reconstructed one. Let $\mathbf{c}$ be a row vector containing the shape coefficients of a given face, then:

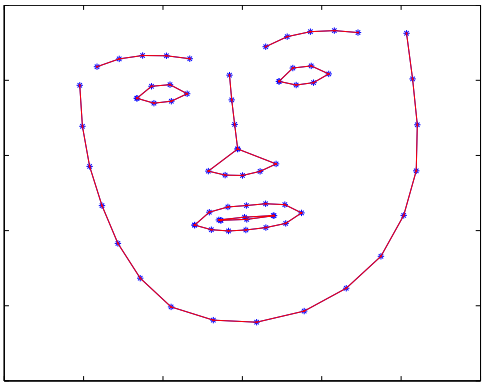
$$\text{error} = \|\text{face} - (\mathbf{c} \cdot \mathbf{V}^{\top} + \text{mean shape}\| \tag{5.6}$$



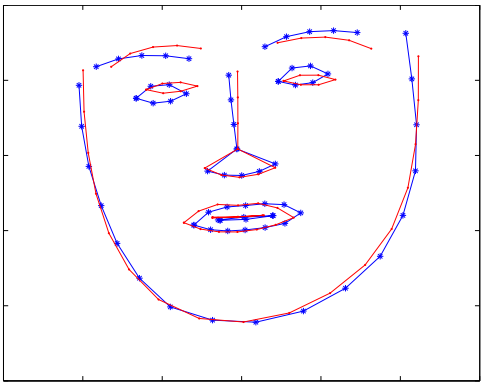
(a) Reconstruction with the whole basis **V**.



(b) Reconstruction with the first 10 vectors of the basis **V**.



(c) Reconstruction with the whole basis **V**.



(d) Reconstruction with the first 10 vectors of the basis **V**.

Figure 5.4: Reconstruction examples. In (a,b): manually labeled data is reconstructed. In (c,d): an unseen face from the same session is reconstructed. Blue: original face. Red: reconstructed face.

5.3.4 Data reduction

The goal of computing the shape coefficients is to use them as features during the classification process. However, having 132 coefficients per frame is a huge amount of data to deal with. To solve this problem we take advantage of the reduction of data the PCA enables to perform. This process consists in keeping the vectors of **V** which provide the 95% of the total energy, and discarding the others. The numbers of vectors needed to keep this percentage of the energy of the faces can be decided by plotting the cumulative sum of the squared elements in the diagonal of the matrix **D** (see Section 3.3.2). As depicted in Figure 5.5, in this particular scenario the number of vectors kept is 10 and consequently, the number of shape coefficients reduces as

well to 10.

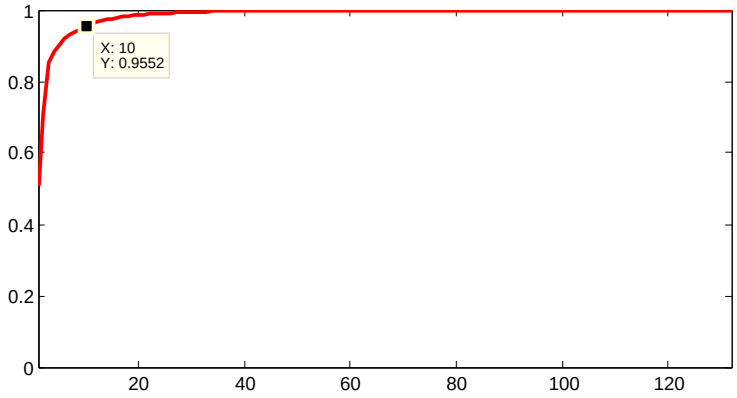


Figure 5.5: Cumulative sum of the squared elements of the matrix **D**, normalized.

Therefore, for computing the shape coefficients of a face, either part of the manually labeled data or simply part of the Spectrum database, only the first 10 columns of **V** are used, resulting in 10 shape coefficients, which will be afterwards applied the three layers of the feature extraction process (frame level, group level, global level) and finally used as features for classification.

5.3.5 Meaning of the shape coefficients

The shape coefficients represent how much a certain vector of the basis **V** contributes to the representation of a given face. At the same time, each one of these vectors represent a deformation of a certain region of the face (in this case, a deformation of a set of landmark points) with respect to the mean face of the faces used for building the model. This deformation results in a movement of the face, such as raising the eyebrows, smiling, etc. The deformation caused by the 10 first vectors of **V**, which are the ones kept after the data reduction process, are shown in Figures 5.7(a) and 5.7(b).

Considering the three angles of head motion as being the pitch, the yaw and the roll (see Figure 5.6), and carefully analyzing Figure 5.7, it is clear that the contribution of the first and second vectors are the pitch and the yaw, respectively. This fact contradicts what was explained in Section 4.4, where it was stated that the aim of the alignment process was, besides the normalization purposes, removing the head motion components of the movement of the landmark points. Hence, it has to be said the the alignment method chosen (Procrustes analysis or similarity

transformation) does not separate entirely the rigid and non-rigid deformations. Making bigger efforts to remove the head motion is part of the future work of this project.

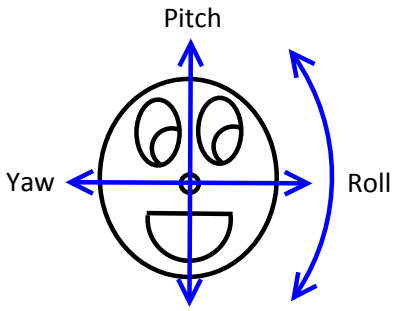
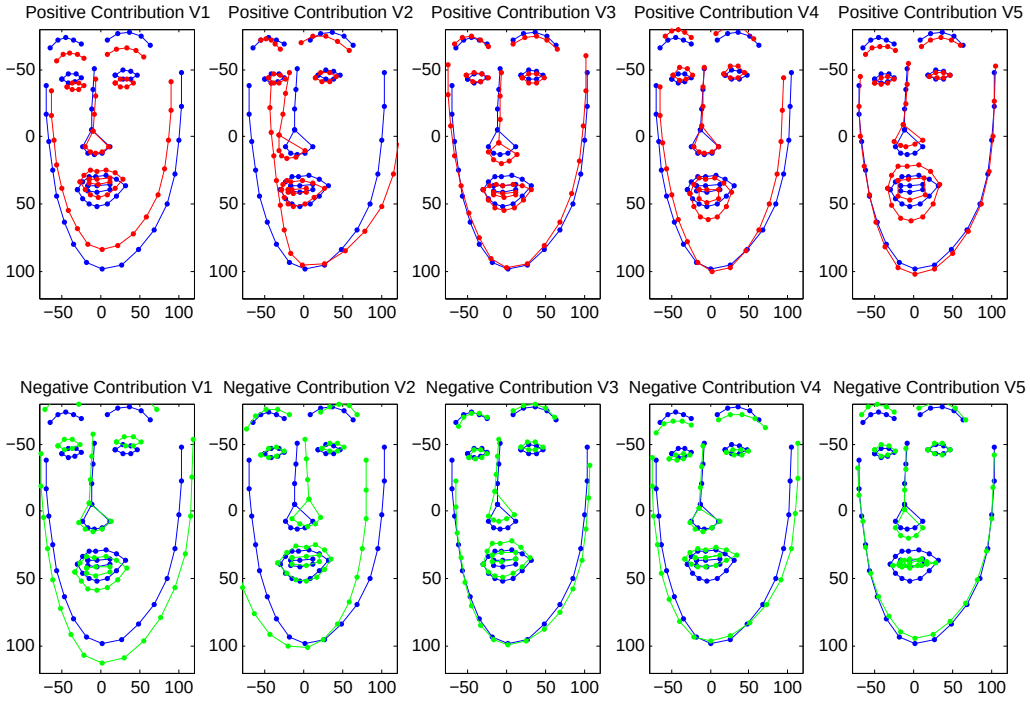
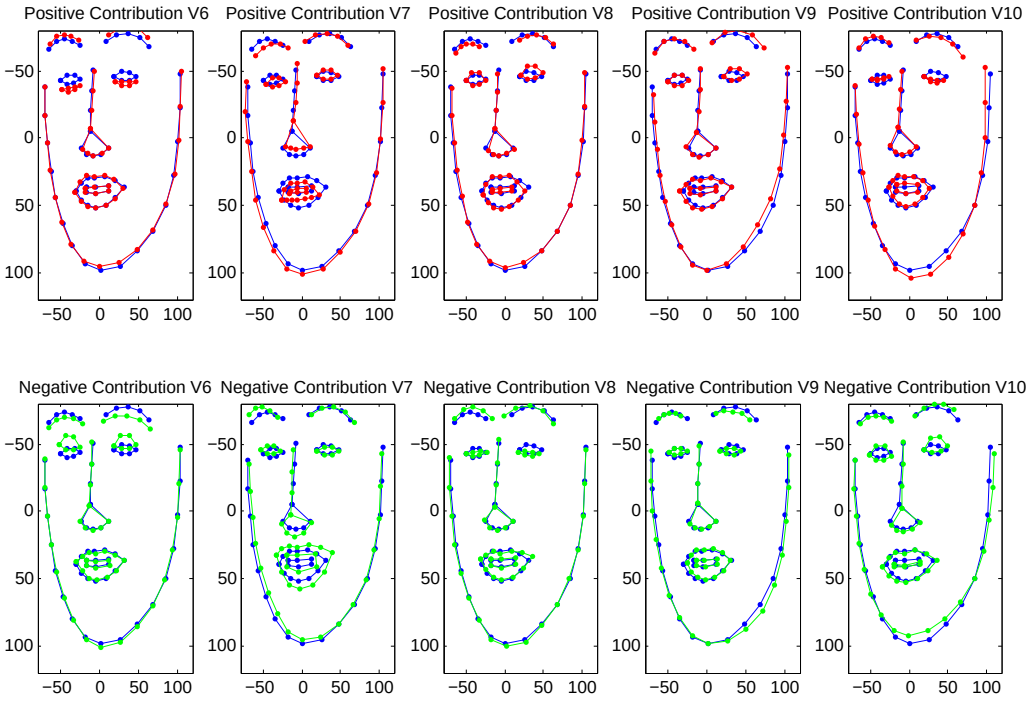


Figure 5.6: Angles of the head motion: pitch, yaw and roll.



(a) Vectors from 1 to 5.



(b) Vectors from 6 to 10.

Figure 5.7: Contribution of the 10 vectors kept after the PCA. Blue: mean shape of the manually labeled data used for building the model. Red: positive contribution (positive shape coefficient). Green: negative contribution (negative shape coefficient).