Automatic Face and Gestual Recognition for Video Indexing

E. Clergue, M. Goldberg, N. Madrane, B. Merialdo
Multimedia Department
EURECOM Institute
2229, Route des Crêtes, B.P. 193
06904 Sophia Antipolis Cedex France

Abstract

Technology is making possible the creation and handling of multimedia documents which contain video segments. At present there is a lack of tools for automatically indexing and representing these documents. In this paper we restrict our attention to a special class of video documents, those sequences containing images of people in movement. These sequences are analyzed and Spatio-Temporal Indices are extracted and used to describe the body motion. The analysis makes use of a 3-D human model composed of an articulated 10 segment structure. The next step is to recognize the identity of the human in motion. To do this we have developed a face recognition procedure based upon projective invariants. Examples are presented and demonstrate the feasibility of our approach.

1 Multimedia Indexing

With the development of Multimedia technologies, we start observing a proliferation of multimedia documents, including a combination of video sequences, audio tracks and text segments. The use of these documents is very attractive, because of the communicative power of media such as audio and video. However, when considered as binary objects that reside in a computer, these documents have the drawback that only a limited number of operations can be applied to them: basically record and playback, with some variants such as fast-forward and backward, compression-decompression, and manual editing such as copy-cut-paste. This prevents any intelligent processing to be performed automatically on these documents. For example, if we store a movie in a multimedia database, and want to retrieve the sequence where the main actor performs a given action, we have to go through a lengthy manual search, eventually browsing the whole document.

The objective of multimedia indexing is to design procedures that are able to automatically define the content of a multimedia document, for example to describe in a movie the times where given characters appear, perform certain movements, pronounce certain words. The techniques for multimedia indexing are based on the combined usage of image, speech and langage analysis and recognition. With multimedia indexing, multimedia documents are no longer dumb binary objects, but acquire a semantic content that can be used to perform intelligent operations, such as automatic indexing, filtering, selection etc...

Along this line, the Multimedia Communications Department at the Eurecom Institute has started a research effort to establish a global environment combining image, speech and language analysis techniques to develop, experiment and evaluate multimedia indexing techniques. This effort is currently focused on three topics: word spotting in an audio flow, body movement recognition and face identification. The present paper describes some of the results that we have obtained in the latter two areas.

2 Body movement recognition

Research in computer vision and scene analysis fields has shown the difficulty of the automatic annotation of video documents. Video data is inherently uninterpreted information in the sense that there currently are no general computational mechanisms for content-searching video data with the precision and semantic exactness of generalized textual search. When imposing specific restrictions onto the scene, the complexity of the problem can be reduced. One such restriction is the analysis of human body motion in a monocular sequence. This scenario is relevant for many applications.

Even though research on human motion analysis and machine vision has become very active in the last few years, few papers have been concerned with non-rigid motion [1],[4]. Interesting analyses of human body motion can be seen in papers of the cognitive science field. Johannson shows an example of interpretation of motion with small lights attached to a human body at several points [2]. O'Rourke and Badler [3] analyzed human motion using constraint propagation and a detailed model of the human body. Nevertheless, these analyzes are very difficult if the person is surrounded by many objects as is common in a real scene.

In this contribution, an approach to automatically analyze human body motion from image sequences is presented. *Spatio-Temporal Indices* are extracted and describe the scene in terms of basic events. This is achieved by a model-based dynamic scene analysis combined with a high-level search algorithm.

2.1 3D human model

The 3D human model we use is composed of an articulated structure and a surface characterization. As shown in Figure 1, the model is based on 10 rigid components whose relative position is described by articulation constraints. Each component is composed of a closed 3D mesh of planar surface patches that approximate the surface shape. We can define the body's position and configuration in space by specifying its reference point and the orientation of each component in the articulated structure.

Concerning the articulation constraints (spatial constraints), we limit the amplitude of the 3 rota-

Figure 1: 3D articulated human model

tions around each articulation. In our analysis, we ignore the accuracy of the 3D structure of the individual components and instead exploit the articulation constraints.

Regarding the surface characterization, each surface patch of the 3D structure is associated with a vector of charateristic parameters. A connected set of surface patch with the same parameters is called a 3D region. In other words, we constitute a mapping of 3D regions on the model. Each component is composed by one or more 3D regions and each region is attached to a set of characteristic parameters.

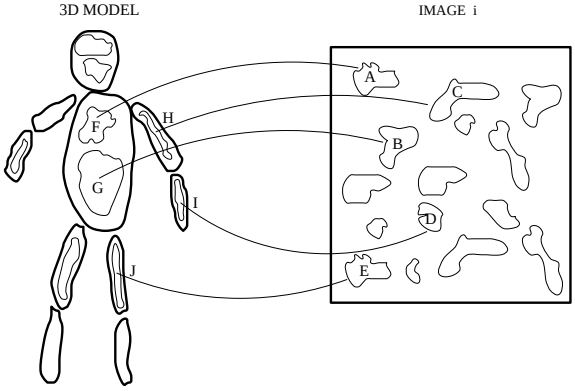

Figure 2: A Generic Hypothesis

2.2 Generic Hypothesis

To work at a high level of abstraction, we must define the basic entity we use for the dynamic scene analysis : the *Generic Hypothesis*. It is represented by a simple matching between a subset of 2D regions located in a particular frame and a subset of 3D regions mapped on the 3D human model. For example, the *Generic Hypothesis* represented in Figure 5 expresses the fact that the regions A, B, C, D, E are the projection of the 3D regions F, G, H, I and J respectively. This definition permits a model-based high level analysis while preserving a simple link with the extracted primitives. To avoid problems of splitting-merging regions we allow a one-to-many mapping between 3D regions and 2D regions. We also allow unlabelled mapping to match an unknown region. In Figure 6, a *Generic Hypothesis* concerning the head is represented. The 3D region of the face is matched with two different 2D regions and there is one unlabelled region.

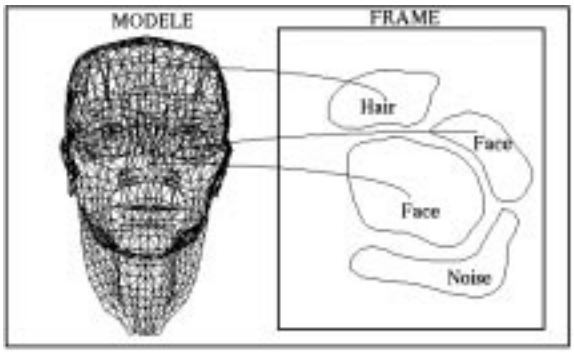

Figure 3: A Generic Hypothesis concerning the head

2.3 Analysis

We exploit scene-specific knowledge to guide the analysis process. This knowledge is formulated as a set of specific hypotheses that are verified throughout the analysis.

The analysis of a raw video document is achieved by integrating six main steps :

1. Initialisation of *Generic Hypotheses*.
2. Creation of the *Search Space*.
3. Search of *Micro-Scenarios*.
4. Identification of a *Macro-Scenario*.
5. Post-processing of the *Macro-Scenario*.
6. Deducing *Spatio-Temporal Indices*.

Given one of the ten rigid components of the 3D human model, these steps are applied successively. This process yields a number on candidate interpretations of the motion of this component.

The first and second steps concern the initialisation phase. A set of good and plausible *Generic Hypotheses* is created and an appropriate *Search Space* is built. The third and fourth steps concern the model-based high level dynamic scene analysis. The spatial constraints of the 3D human model are used in conjunction with a search heuristic in order to identify a scenario describing the whole sequence. In the fifth step, we try to correct segmentation errors and improve the accuracy of the scenario. The last step concern the description of the sequence at a high level of abstraction.

2.4 Initialisation of Generic Hypotheses

In the first step, we create a set of *Generic Hypotheses* for a particular frame of the video, based on the primitives previously extracted in some way. Using *a priori* information about the structure of the objects in the scene, we reduce the number of potential *Generic Hypotheses*, selecting and retaining only the most coherent. More precisely, we define a coherence criterion to filter all the most likely *Generic Hypotheses*. This criterion depends on the relative position of the 2D and 3D regions and on the similarity of the characteristic parameters between the 2D and 3D regions.

2.5 Creation of the Search Space

The set of *Generic Hypotheses* created in the first step is not structured and therefore is not suitable for a high level analysis. In other terms, we must introduce a relation between two *Generic Hypotheses* in the same frame or in two consecutives frames. The second step allows to create such a hierarchy and store the *Generic Hypotheses* in a *Search Space*. This *Search Space* is a 3D graph as shown in Figure 8. Each plane represents one frame and the nodes are the *Generic Hypotheses*.

We make the distinction between two kinds of links : the intra-frames links and the inter-frames links. The intra-frames links represent a relation between two *Generic Hypotheses* in the same frame and symbolize a spatial constraint as defined by the relation $H_1 \subset H_2$. The inter-frames links represent a relation between two *Generic Hypotheses* in two consecutive frames and thus symbolize a motion constraint.

The creation of the links is one of the most important steps in the analysis. To reduce the complexity of the search, we need to limit the number of links. This problem is adressed in the next step.

2.6 Creation of the Micro-Scenarios

A *Micro-Scenario* is a short time succession of *Generic Hypotheses* and is represented by a small path of inter-frames links as shown in Figure 9. Because of its short lenght, a *Micro-Scenario* is robust and easy to extract. Experiments show that a lenght of 3 or 5 frames is sufficient. We note $First(S)$ and $Last(S)$ the first and the last *Generic Hypotheses* of a *Micro-Scenario* S. $Frame(H)$ denotes the frame number associated with a *Generic Hypothesis* H.

Micro-Scenarios are created by simple matching of adjacent *Generic Hypotheses* in time domain. Two consecutive *Generic Hypotheses* are linked and integrated in the same *Micro-Scenario* if they are spatially close.

The set of *Micro-Scenarios* is sparse and the *Search Space* comprise a number of discontinuities or areas without *Micro-Scenarios* called *Holes*. We distinguish two kinds of *Holes* :

- A *First Order Hole* is a *Hole* between two *Micro-Scenarios* S_1 and S_2 with $Frame(Last(S_1)) = Frame(First(S_2))$. In other words, the last *Generic Hypothesis* of S_1 is in the same frame than the first *Generic Hypothesis* of S_2. An example is given in Figure 10.

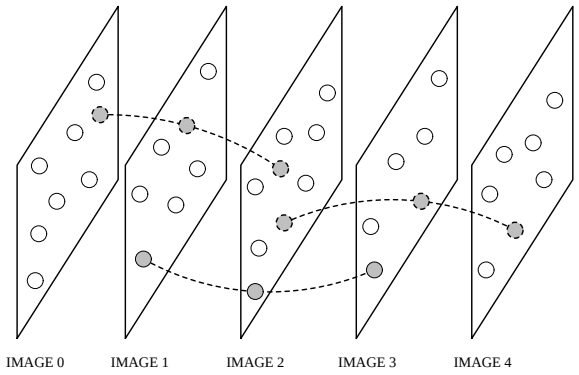

Figure 4: Some Micro-Scenarios

- A *Second Order Hole* is a *Hole* between two *Micro-Scenarios* S_1 and S_2 with $Frame(Last(S_1)) < Frame(First(S_2))$. An example is given in Figure 11.

In brief, a *Micro-Scenario* expresses a local and robust knowledge in the sequence as "Something here looks like the head and is moving from this point to this point between frames 10 and 13" for example. It is composed of a spatial knowledge (*Generic Hypotheses* concerning the head) and a temporal knowledge (motion of the head between frames 10 and 13).

2.7 Identification of a Macro Scenario

A *Macro-Scenario* is a long and complete time succession of *Generic Hypotheses*, describing a video cut in terms of consecutive events. It is represented by a long path of inter-frames links.

We grapple with the problem of the identification of a *Macro-Scenario* by navigation in the *Search Space*, searching the most coherent path from the first to the last frame. The idea is to find a global path from the first to the last image of a video cut which go down through the sparse set of *Micro-Scenarios* in a coherent way. Each time we go down a *Micro-Scenario* a pseudo-distance is used to determine the next *Micro-Scenario* to use. Then we solve the *Hole* problem to make a jump between the two *Micro-Scenarios*.

Solving the *Hole problem* for a given pair (S_1, S_2) of *Micro-Scenarios* depends on the nature of the *Hole*. For a *First Order Hole* a spatial interpolation is used. The idea is to create one or several intermediate *Generic Hypotheses* to make a link between the last *Generic Hypothesis* of S_1 and the first *Generic Hypothesis* of S_2.

For a *Second Order Hole* the problem is much more difficult since a jump must be created in both space and time domain. Two approaches are proposed :

- In the first approach, we delete most of *Second Order Holes* by increasing the number of *Micro-Scenarios*. The problem is that we also increase the complexity of the search.
- In the second approach, a special link is created from $Last(S_1)$ to $First(S_2)$ and correspond to a search failure. This search failure occurs for

example in the case of a large occlusion of a component. We are investigating solutions to this failure problem.

2.8 Future work

Experiments have demonstrated a satisfactory accuracy and robustness regarding some occlusions and noise. We shall now consider other important aspects of future work, including the development of a complete computer tool for video indexing. In particular, we are investigating various approaches to improve reliability and robustness.

3 Face Recognition

Face recognition is a task that human beings accomplish really easily every day, even in the worst conditions like aging or insufficient illumination ... For many years, this task which seems so natural for us, has become a real challenge in computer vision.

Among the information that we would like to get about a video sequence is the identification the person(s) that appear on this sequence. While identification should take into account all features of the individual (face, body, attitude, movements etc...), it appears that face recognition is a very important component in such a mechanism. The global question of face recognition can be decomposed into a number of sub-problems: - detection of faces in a video (presence and location), - person recognition (by comparison with a database of known individuals), - person discrimination (comparing if two images represent the same person or not).

All these themes are important in various applications scenarios. For example, person recognition can be used when we want to have a filtering agent that watches TV news in our place and records all appearances of given people (the president of a foreign country if we are interested in foreign affairs). Person discrimination is useful when we want to automatically create a "table of contents" of a movie, showing all persons who appeared in the movie (but without necessarily knowing who they are).

The video sequences we work on, can be movies, or TV journal or filmed meetings. In these conditions, it is easy to imagine that we have no control on the restriction of the orientation of the face and its location, illumination, distance to the camera and facial expressions. Consequently, most classical methods can not be directly applied. The algorithm we need, has to be able to differentiate or to recognize several persons as well.

Most Classical methods require specific conditions when acquiring images. For example, technics based on characteristics measurements (eyes, nose, mouth ...) or recognition from profile views need to fix the distance to the camera, the orientation of the face, the illumination ... [6], [7], [8].

For some other approaches like neural nets or eigenfaces[9], you need to take a lot of pictures for different chosen values of orientation, illumination and distance to the camera, that is not possible for us to constraint and control.

These restrictions are necessary for the robustness to the noise, and to avoid the effect of any perspective transformation.

3.1 Recognition based on invariants

Our first approach is to take significant measures on the face. But we have no control on the location and orientation of people in the image.

From the work of gesture recognition presented in the previous chapter, we obtain the localization of the head. At this step of implementation, the required characteristic points are manually selected.

To be as independent as possible from the perspective transformations, two projective invariants have been defined on faces for the first two identification parameters.

As a first characteristic, we decided to adapt an approach which is commonly used in pattern recognition: projective invariants. We tried to define a cross ratio from four points on a face. To calculate it, these points have to be aligned, and then the cross ration is independent of any perspective transformation (combination of 3D rotations, 3D translations and zooming).

For four points A, B, C and D, the associated cross ratio is given by the following formula:

$$R(A,B,C,D) = (d(A,C)*d(B,D))/(d(A,D)*d(B,C))$$

The set of points has to be enumerate in the same order from one view to another.

The first set of points is given by the corners of the eyes see figure (5). You can see the result of such a measure on the figure (7) for five persons in five different attitudes and expressions. You can remark that for each person (subset of five successive measurements) the calculated invariant values are roughly independent. The existing differences come from the manual selection of the points.

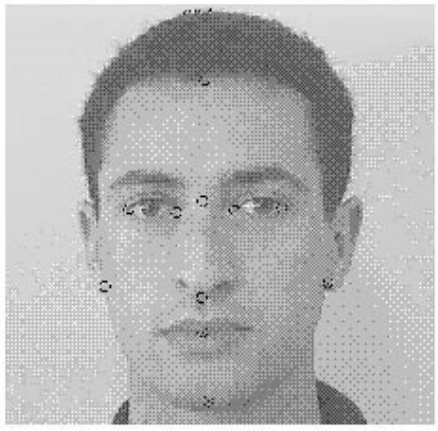

Figure 5: This figure show the characteristic points required for the invariants and the orientation parameters calculation. At this time, they are manually selected.

Of course this information is not sufficient. We can define another cross ratio with the top of the front, the top and the base of the nose and the center of the mouth.

3.2 Recognition using face orientation

The invariant based on the eyes is not always defined according to the orientation of the subject. As a consequence, we need this pose information to be able to cancel this measure if necessary. Using a generic 3D model of a human head and selecting few points on the face in the first frame, we can obtain a rough estimate of the initial pose and track these points over the sequence.

To do that we used the Powell's quadratically convergent method to minimize the nine parameters describing the three 3D rotations around the symmetry axe of the face, the three translations and the three zooming factors. The function to minimize represent the quadratical errors sum of selected points location in the image and the projection into the image of their 3D correspondants once the perspective transformation has been applied.

This information will be of a good help for example to select over a set of successive frames the "best" orientation, that is the nearest from the front view.

From the 3D orientation information, we can define the 3D symmetric axe S of the face. We can then define different 3D planes like the one passing through the corners of the eyes and perpendicular to S. We can take then into account for the recognition if this plane cuts the ears. The same kind of plane can be defined by the base of the nose, and look again for the intersections with the ears. You can build other planes parallel to S passing through the corners of the mouth or the sides of the nose and check if they cut the eyes. These binary information turn out to be significant to differenciate people.

These measurements allow us to make a first classification of the persons appearing in the video sequences.

In order to validate these measurements, we are building a hierarchical classifier. The aim of this classifier is to combine the unhomogeneous caracteristics in view to differentiate or to recognize faces. By statistical approach, we will obtain a normalization of the different identification parameters. Then it will be possible to compare different faces.

3.3 Future work

As a first step, we want to automatically detect the set of characteristic points we need to calculate invariants and orientation of the face.

We have color information from video images and we want to use it simply for the recognition process. From the RGB information we take the chromatic value of the hair and the eyes by projecting on the chrominance plane. This projection aim to be roughly independent of the illumination. Then we have 2D vectors of information which are easier to compare in terms of color regions for the classification and the recognition.

Our future approach wille be to construct a hierarchical classification of the different parts of a face which have a meaning role on the identification process. This kind of method will lead to a sort of identikit picture. To describe someone you will finally give the list of classes, corresponding to the successive characteristics or meaning parts of the face. This approach will lead to a significant gain in time processing and memory space used (the description is contained in a vector, some classes will be common for several persons ...).

4 Conclusion

In this paper we have presented a tool for the automatic indexing of multimedia documents containing video sequences of humans in motion. We describe an analysis technique which allows to detect and then describe humans in motion using a 3-D articulated 10 segment model. The result of the analysis are Spatio-Temporal Indices. We then show how the identity of the human in motion can be deduced by using projective invariants. Our work is the first step to a more complete video indexing system which will also be able to handle speech.

References

[1] Huang, T.S., "Modeling, Analysis, and Visualisation of Nonrigid Object Motion" Proc. of International Conf. on Pattern Recognition, Vol. 1, pp. 361-364, Atlantic City, NJ, June 1990.

[2] Johannson, G., "Perception of Motion and Changing Form", Scandinavian J. Psychology, Vol. 5, pp. 181-208, 1964.

[3] O'Rourke, J. and N.L. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 6, pp. 522-536, 1980.

[4] A. Shio and J. Sklansky, "Segmentation of People in Motion", Proc. IEEE, vol. , no. , pp. 325-332, 1991.

[6] J.Y. Cartoux, J.T. Lapreste, M. Richetin,"Face authentification or rexcognition by profile extraction from range images", IEEE Workshop on Interpretation of 3D scenes, Austin, Texas, November 26-29, 1989.

[7] M. A. Shackleton and W.J. Welsh,"Classification of Facial Features for Recognition", Proc.,p 573-579, CVPR, Mauï Hawaï, June3-6, 1991.

[8] A.L.Yuille, D.S.Cohen and P. W. Hallinan,"Feature Extraction from Faces Using deformable Templates", Proc CVPR, San Diego, Californie, June, 1989.

[9] M. A. Turk and A. P. Pentland,"Face Recognition Using Eigenfaces", Proc. CVPR, Mauï, Hawaï, June3-6, 1991.

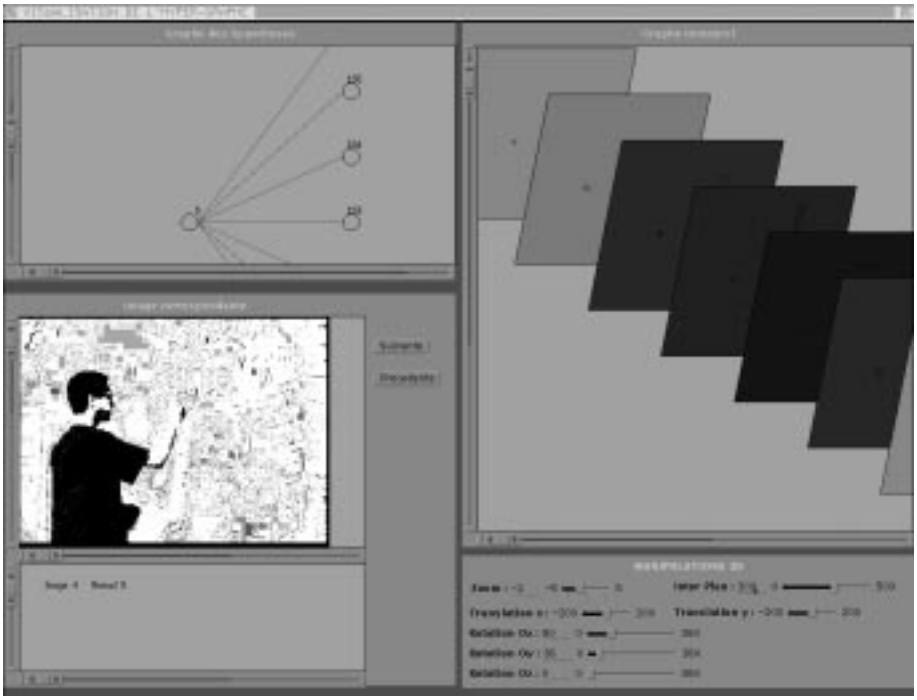

Figure 6: Graphic Interface

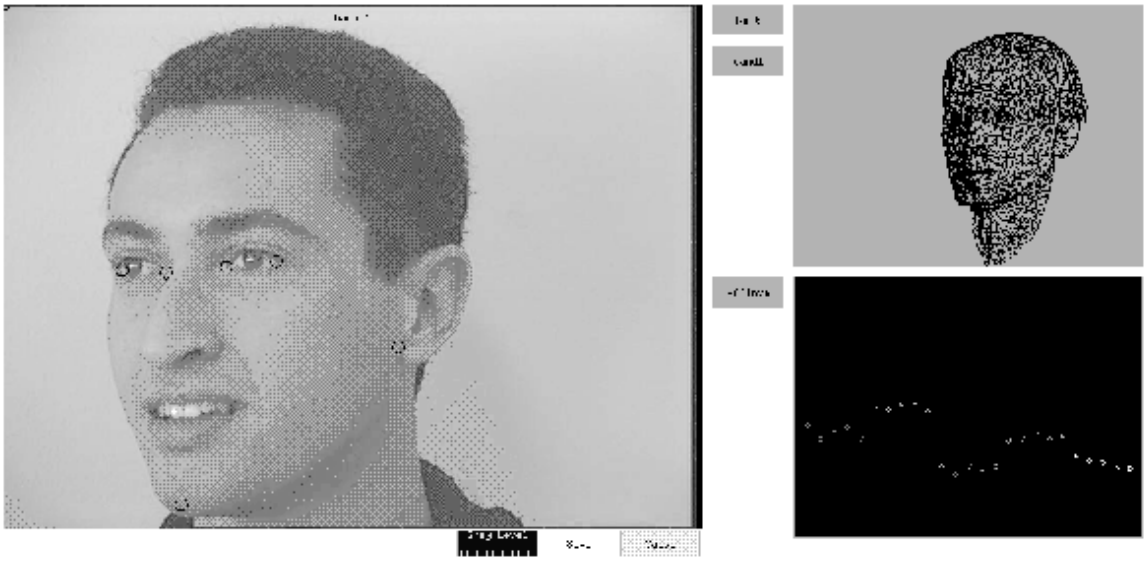

Figure 7: This is a part of the interface. In the main frame, you can see the initial image where some points have been selected for the invariants and the orientation. You can remark that once the orientation parameters are known, they are applied to the 3D model which is projected onto the image. This simple minimization gives a good estimation of the pose. In the upper right window, you can see the representation of the 3D model oriented with the parameters obteined by the minimization. In the last part, these are the values of the invariant based on the eyes. You can see 25 values. In fact, there is five series corresponding to five different persons. Five images for each of them in different pose and expression have been studied. The rough invariancy of the value over each subset can be remarked.