Human Gesture Segmentation Applied to Dance Performances

Frédéric LEAU
Projet de Fin d'Etudes
ENSEA, Cergy-Pontoise section STC

IRCAM, Paris

Departments involved: Production-Research
Teams: Applications Temps-Réel / Pôle Spectacle
Supervisers : Emmanuel Fléty-Frédéric Bévilacqua-Rémy Muller

February / July 2005

Contents

List of Figures

Acknowledgments

I would like to thank Emmanuel Fléty who gave me the opportunity to work at Ircam during the last five months, helped me with Rémy Muller and Frédéric Bevilacqua in the task of writing my report/presentation, and the whole team Applications Temps Réel for their welcome.

Introduction

IRCAM -Institut de Recherche et Coordination Acoustique/Musique- was created by Georges Pompidou in 1969 and was directed by the famous Pierre Boulez. It was and is still dedicated to contemporary musical research. Since, Ircam has always tried to keep the deepest link between artistic performances and scientific research activities. In this context, a new research group has been formed at IRCAM since June 2003 and aims at developing new technological tools for artistic performances. This group called -Performing Arts Technology Research Team- is in charge of transversal projects and involves people from different departments of IRCAM. These people are members of the Real-Time Applications Team (Research department) and the production department, and are working together on both human gesture capture and analysis. More and more artists give rise to this research domain because it could allow them to create new kinds of performance combining different forms of arts such as dance, music, theater, circus... In such a way, they can create new interactive installations capable for example of involving the audience and many more new interactive processes.

Two main approaches are being considered at Ircam : in the one hand, the analysis of the instrumental gesture of a violonist in real time and in the other hand the analysis of dance performances which main application would be the tracking of the dancer's movement so that the dancer could be able to trigger musical, video, or visual processes.

During my internship, my work has been focused on the human gesture analysis applied to dance performances and consists of the dancer's movement segmentation.

In the first part, this document will present the main requirements, including the state of the art concerning my subject. Then I will detail the different approaches developped at IRCAM in this field. Next, I will highlight the context of my study with the main tools I will have used during my internship.
In the second part, the different methods I implemented will be presented. Finally I will try to focus on the main results concerning the data segmentation that I have been able to achieve.

Part I

First Requirements

In this part, I expose the terms of my internship and try to highlight some contextual aspects in regard of what have been done previously at IRCAM and in the different scientific communities.

Chapter 1

State of the Art

I spent my first month at Ircam dealing with the establishment of the state of the art. I started it with the different statements of my subject which can be summarized with the next sentence :

segmentation and decomposition of the movement using computer vision and sensor-based techniques in the context of contemporary dances

This implies several concepts to be defined or at least explained :

- First the segmentation and decomposition are typically signal processing techniques that come into a more general framework : the analysis of the movement.

- The kind of movement needs to be defined and/or detailed. In this manner, the context can bring precise informations on it.

- The different technologies involved in the gesture captation system and the different methods used to perform the tasks listed above.

- Finally the context is very important because it will define the choices, constraints and applications of the system to be developped.

1.1 The human gesture analysis

Human gesture analysis is getting more and more interest these days according to the increasing number of articles relating to this wide domain of research. This can be explained for different reasons:

- the improvment of computer performances which are more efficient to answer to the memory capacity and CPU power required by most of the algorithms used in this domain.

-the growing number of applications using this kind of technology that find benefits in such a research effort.

The human gesture analysis is in keeping with the general analysis of the movement of objects. In general, the motion of physical objects is non-rigid [ACLS94], i.e. distances and angles of the different parts of the object are not preserved. The movement of the human-body is non-rigid but also articulated. This means that every part of the human body has a rigid or quasi-rigid movement but the whole human body has a non rigid movement.

Depending on the application context, different tasks are being performed in the general analysis framework of the movement. These are the recognition, classification, detection, modeling, segmentation, decomposition, representation, feature extraction, synthesis, following of human gestures (see the articles that establish a review [Tur] [Gav99] [JKA99]). In the representation process, the way the human body is modeled also depends on the application. These representations can be stick figures-skeletons (based on points 1D), blobs (2D), 3D-volumes. Besides, most of the applications I have seen try to deal with only a part of the human body. In general, these are the face, the hands or both. In the last case, the approach is considered as multimodal. Only a few articles relate to the process capable of handling the full human body as it is a constraint in our context [SDP00].

1.2 A gesture: a context dependant definition

1.2.1 The context of contemporary dance

Unlike the music which can be stored on a cd or can be written with a score, contemporary dance has no comparable media to save and write its choreographical performances. With the camera, it has become possible to record dances but there is not really a way to write choreographies. Rudolph von Laban has established a way to write choreographies [Rin]. This notation is nowadays the most commonly used in the domain of contemporary dance but still, the choreographic langage is so subjective and inherent to its composer that it remains very difficult to write choreographies. The main cause is that it is particularily difficult to define unit gestures in a contemporary dance. More classical dances such as ballet can easily be written due to the fact that it relies on a set of defined movements [SEY$^+$], a dictionnary of unit gestures. The principle of contemporary dance is a little more abstract. Movement is perceived as a flow that goes freely from one body part to an other part and so on. No constraints are defined, every choreography is dependant on the vision of its choreographer [DB]. Therefore a major constraint to create a general gesture following system in that context is that it has to fit to any choreographical langage without having a precise definition of a gesture.

1.2.2 A model of gesture

According to the context described above, there is no definition of a unit gesture. Therefore, it is particularily difficult to model a gesture. Anyway, some researchers try to match a model with their interpretation of gestures which are once again context dependant. Most of the time, these models are based on a multi-layer architecture.
Aaron Bobick (MIT Chicago [Bob]) decomposes the human motion in three different levels:

- the *movement* or atomic gesture. It corresponds to a static gesture or posture.

- the *activity* is a sequence of *movements*. This layer is in general handled with Hidden Markov Models or Dynamic Time Warping.

- the *action* which is a movement that people usually describe when they are doing something.

1.3 The commonly used techniques

1.3.1 The hardware technologies

To perceive the movement and especially the human gestures, different hardware configurations are actually being used in some of the research communities.

-Motion Capture

Motion Capture refers to the computer hardware and software that makes possible recorded digital 3-D representation of moving bodies. Recording sessions involves the placement of markers or sensors on strategic positions on the body that provide the basic information for the computer software. The expense of these systems, which includes the cost of the equipment as well as the expertise to run it, is enormous with developments being driven primarily by those industries such as medical, military, entertainment and advertising that have the necessary capital. Compared to the other gesture acquisition systems, this technology gives results that are very close to the reality. With this, problematics such as the study of morphological aspects of the human body can be performed

[KTP03] [Tro02].

-Traditionnal computer-vision based systems

These systems differ from the motion capture by the cost first and then by the fact that most of the time they cannot manage and solve the occlusion problems. Several systems exist. They deal with a different number of cameras. Depending on this number of cameras, they establish different models of the human body. Most of these are working in real-time.

- Sensor-based systems

The motion is captured with some sensors fixed on the human body. With this technology, a representation of the human body is not possible. The main goal is to get informations of body parts such as the position, the acceleration, the flexion etc..
Paradiso [BP02] has used this kind of technology. Several sensors are disposed in a cube which can represent until six degrees of freedom. Most of the time this technology is added to an other gesture acquisition system to get a full set of parameters that can be furtherly derived to get efficient motion descriptors.

1.3.2 The signal processing techniques

In the domain of computer-vision, algorithms such as the optical flow or the color-tracking [BD02] one are used to perform the motion tracking in the field of view of the camera.

According to the definition of gestures that can be either static or dynamic, the use of machine-learning methods is widely spread in the human gesture analysis. Indeed, the temporal aspect of a dynamic gesture is so complex that it seems to be almost random. This can only be handled with the use of machine learning tools such as Hidden Markov Models, Bayesian Belief Networks, Neural Networks, Finite State Machines etc..[HJLA01] [Rab89]. The main principle is quite simple, a database is established to serve as an example for these machines. These machines learn/train according to these examples. With the use of a similar database (the test database), it is possible to check if the machines have been well-trained or not.
In fact using these techniques could seem to be an easier way to achieve a segmentation for example. But the difficulties still remain. They consist here in the configuration of these machines which is not trivial.

Chapter 2

IRCAM's involvement in human gesture analysis

Ircam is involved in the human gesture analysis for at least five years now. A new group has even been created: Performing Arts Technology Research Team. This group is in charge of this research and its applications to live performances such as dance, theater, music.
We can divide the work done on this research domain in three main parts:

- the development of motion and gesture capture systems in regard of the constraints of performing arts.

- the human gesture analysis using the data collected.

- the integration of these technologies in artistic projects

2.1 The two gesture acquisition systems used at IRCAM

To analyse the human gesture data, we need first of all an interface capable of converting human gesture into digital data and a computer to store and analyze it.
At IRCAM, two acquisition-systems are actually being used, including one developped by Emmanuel Fléty and Nicolas Leroy:

2.1.1 Ethersense/Wisebox

The Ethersense is a versatile sensor acquisition system based on network technology which has been developped at Ircam by Emmanuel Fléty and Nicolas Leroy [FB04]. It is an interface which allows the analog sensor's data to be converted into digital data every 1-2 ms and transmitted at a very high rate to the computer using a 10 Megabits ethernet communication layer. This conversion is held by a new protocol called Open Sound Control . This protocol is dedicated to the communication among computers, sound synthesizers, and other multimedia devices and is particularily adapted to gesture acquisition due to the fact that it is easy to plug-in, has a low latency.

Figure 2.1: The Ethersense interface

The main difference between Ethersense and the Wisebox is that the Wisebox is the WIFI version of the Ethersense interface. Its main advantage is the fact that it is wireless. Therefore, it is particularily adapted to dance performances because dancers do not have to deal with wires between the computer analysing/storing the data in real-time and the interface connected to the sensors.

Figure 2.2: The Wisebox interface

2.1.2 Eyesweb

Eyesweb is a software environment dedicated to video processing. Eyesweb and its libraries of patches are the result of a research work concerning algorithms

and computational models for real-time analysis of expressive gestures in human full-body movement.
This work has been carried out at the INFO-MUS Lab (Genova, Italy [eyw04]) in the context of the MEGA Multisensory Expressive Gesture Applications and is focused on the analysis of expressiveness in human gestures.
At its origin, Eyesweb has been conceived for the design and development of real-time dance, music, and multimedia applications. Many libraries have been developped and some especially for motion analysis such as the Eyesweb Motion Analysis Library. This library is in fact a collection of modules for real-time motion tracking and extraction of movement cues from human full-body motion via a monocular videocamera.
Programming with Eyesweb consists of building patches, i.e. connecting objects with cords.

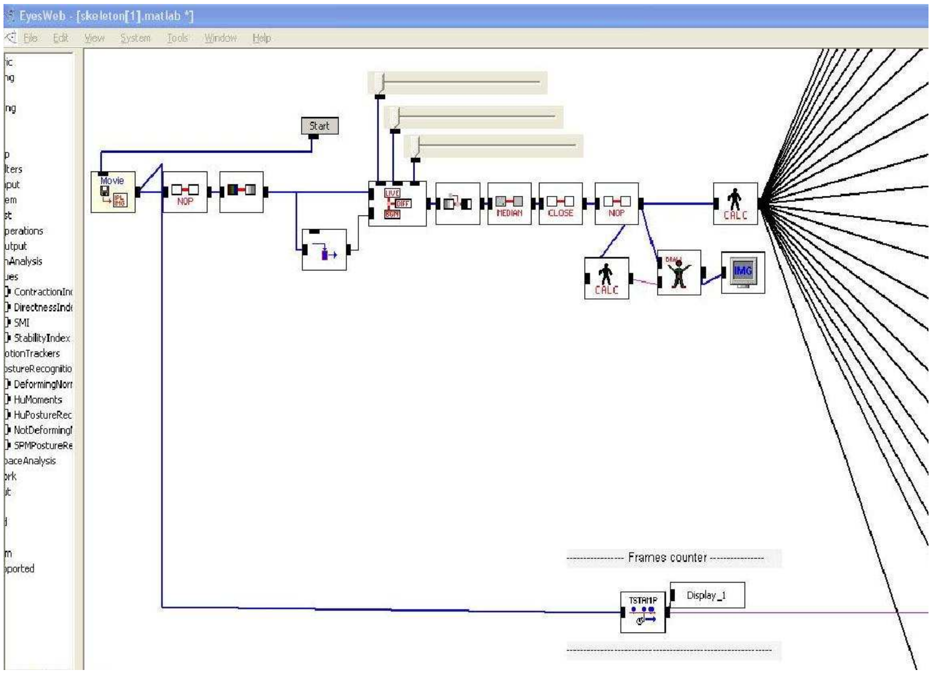


Figure 2.3: An eyesweb patch

Its main application to dance consists of evaluating expressivness, fluency and feelings of the different dancers who are executing choreographies. To that purpose, Camurri and collaborators (INFO-MUS Lab Genova Italy [eyw04]) describe methodologies to measure these expressive cues by applying a multi layered signal processing architecture. This architecture is described more precisely in the next section.

To conclude, the combination of these two gesture capture systems allows us to have a quite large and robust enough set of motion descriptors to perform accurate gesture analysis.

2.2 The general context of the human gesture analysis at Ircam

2.2.1 Different gestures/different analysis

Two main approaches are being considered in this task:

-The augmented violin is a project centered on the use of a regular acoustic violin with added gesture capture technology integrated on the bow (based on the use of the Ethersense technology). The general aim is to provide musicians with novel opportunities for mixed acoustic/electronic music. However this approach doesn't come in the scope of this report.

-The second approach is to develop a general framework for gesture analysis, modeling and recognition, compatible with the constraints of performing arts such as dance. The analysis should provide the ability to follow or recognize human motion and action, and to derive high-level movement parameters, for example related to style ,affect or choreographic langages.

The main difference between these two approaches relies on the fact that they are not dealing with the same kind of gestures. Indeed, gestures of the musician are contextual to the use of his instrument. On the other hand, gestures of the dancers are professionnal movements and gesture acquisition systems should take advantage of their ability to reproduce expertly dance movements.
The next section only refers to the dance context.

2.2.2 Several degrees of interactivity in the dance context

Depending on the artistic project involved and the needs of the choreographer, the use of a gesture capture system can be very different and therefore implies various constraints on the system. These differences can also be seen as the degree of interactivity the choreographers want to have in their pieces.

- The system can be used as a tool which is used during training sessions or during the dance/music creation/writing process. In this situation, real-time use of the system is not required.The degree of interactivity can formally be considered as low.

- The system implies the active participation of the audience which is not considered as a passive spectator anymore. He has a limited control of some events of the piece. In that case, the degree of interactivity is medium.

- Finally, the piece includes strong relationships between the dancers and some musical, and/or graphical processes. The degree of interactivity is then considered as high. This way, real-time performances are required and the system has to be able to perform accurate following of the movements of the dancers.
The Wise Box interface has been designed to enable such an interactivity in real-time.

Depending on this degree of interactivity, two modes can be included to this framework:

-non-realtime: this mode corresponds to the analysis of a gesture database. This mode is used to formalize a gesture vocabulary (context dependant), which will be used for the real-time mode. For example, automatic motion segmentation, gesture labeling can be performed. That is most of the work that I have been doing during my internship.

-real-time (performance): the system outputs different types of parameters:
–high-rate parameters, to control, for example, of continuous parameters in sound processes
–low-rate parameters, corresponding to the recognition of relatively long gesture phrases, which can be used for example, to define a particular context of gesture-sound interaction, like choreography following.

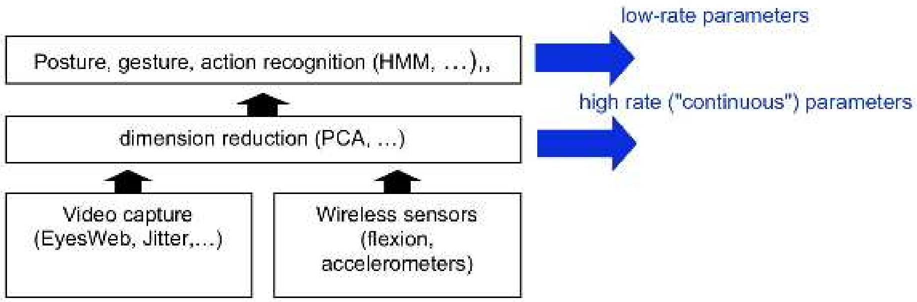

low-rate parameters
Posture, gesture, action recognition (HMM, ...),,
high rate ("continuous") parameters
dimension reduction (PCA, ...)
Video capture (EyesWeb, Jitter,...)
Wireless sensors (flexion, accelerometers)

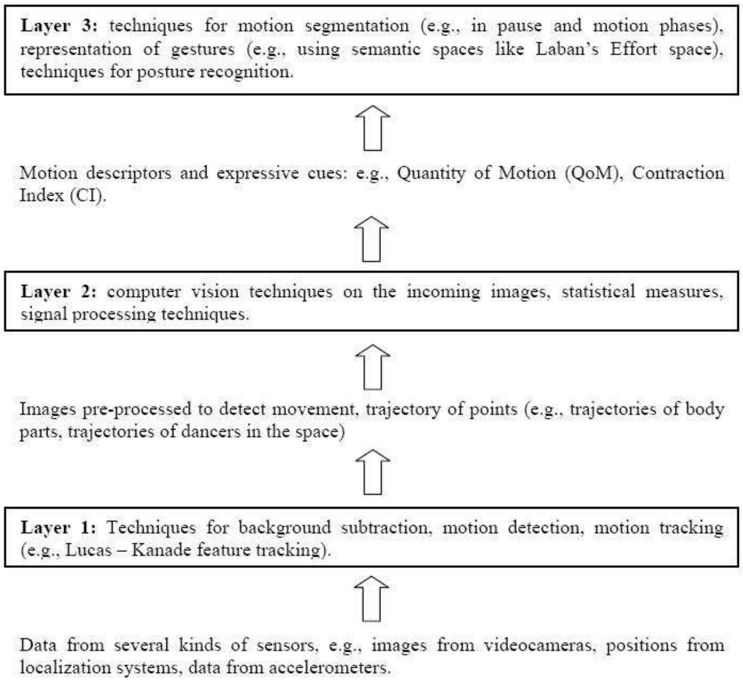
Figure 2.4: Scheme of the real-time system

A previous work done by Rémy Muller [Mul] had for constraints the real-time mode and dealt with the high level of interactivity. It consists of a human motion following system using Eyesweb and Hidden Markov Models.

2.2.3 A multi-layered conceptual framework

The gesture analysis is based on the creation and use of motion descriptors.
A hierarchical description of the motion descriptors has been established in the similar way as it has been done with Eyesweb.

Physical signals coming from the camera or the sensor-based-capture-system such as the position or acceleration of a body part are considered as low-level parameters. In Eyesweb, the techniques used at this level are commonly used computer vision techniques to recognize human motion and activity (background substraction and motion tracking).
Intermediate parameters are obtained by mathematical transformations of the low-level parameters. These parameters can be for example the quantity of motion, position of the center of gravity, stability index, contraction index etc...
High level parameters are those which are close to a choreographic langage. A precise segmentation of the movement can be achieved according to the choreographic context.



Layer 3: techniques for motion segmentation (e.g., in pause and motion phases), representation of gestures (e.g., using semantic spaces like Laban's Effort space), techniques for posture recognition.

Motion descriptors and expressive cues: e.g., Quantity of Motion (QoM), Contraction Index (CI).

Layer 2: computer vision techniques on the incoming images, statistical measures, signal processing techniques.

Images pre-processed to detect movement, trajectory of points (e.g., trajectories of body parts, trajectories of dancers in the space)

Layer 1: Techniques for background subtraction, motion detection, motion tracking (e.g., Lucas – Kanade feature tracking).

Data from several kinds of sensors, e.g., images from videocameras, positions from localization systems, data from accelerometers.

Figure 2.5: The layered extraction of motion parameters

2.3 Artists involved-integration ot the research work

IRCAM is quite unique because there is not so many places where people with scientific, artistic or both backgrounds can work together. As a matter of fact, the team I am working with relies on the support of the artists in residence, choreographers, producers, visual artists... who have to document their reflexions and practices linked to the principal themes within the various areas of research.
Consequently, the team welcomes for the year 2004-2005 the following artists:
Andrea Cera, Hervé Robbe, Nicole and Norbert Corsino, Eric Génovèse, Myriam Gourfink, Patrice Hamel, Emmanuelle Huyhn, Jean-Francois Peyret for the following collaborations:
-With Andrea Cera and Hervé Robbe: Autour de Waves
Andrea Cera, composer, and Hervé Robbe, director of C.C.N. in Le Havre, Haute-Normandie, have initiated a project of musical, audiovisual and choreographic creation. Completion scheduled for summer 2005.
-With Eric Génovèse: Le Privilège des Chemins
Eric Génovèse works at IRCAM on the vocal transformation of actors from la

Comédie Francaise and on sound environments for the production of Privilège des Chemins.
-With Jean-Francois Peyret: Les Variations Darwin
Théâtre Director Jean-Francois Peyret is working at IRCAM in the framework of a commission from the Théâtre National de Chaillot. The project uses automatic text generation set to the music of Alexandros Markeas.

Chapter 3

The contex of my study

3.1 The terms of my internship

3.1.1 Presentation

During my internship, I have to deal with both segmentation and decomposition of human full-body movement in the context of contemporary dance (see section 1). These tasks are part of a general framework described in the previous section. To that purpose, we aim at using computer vision and machine learning techniques . This implies the use of different softwares such as Matlab, Eyesweb, MAX-MSP and hardware such as sensors and videocameras. Consequently, a general background of hypothesis and constraints can be defined according to these hardware and software configurations.

Hypothesis-constraints:

- 3D human motion is analyzed via a 2D vision system.
- The camera is fixed and the focal plan must not change.

- The lighting has to be constant.

- There must only be one dancer in the field of view of the camera.

- The only visible motion has to be the human one over a constant background.

- The size of the performer has to be comparable with the one of the image.

- It is quite difficult to perform real statistical analysis according to the few examples of a choreography that are available.

The following paragraph presents the database I have been using in my work. This database relies on the set of hypothesis listed above. Then we will see how to deal with Eyesweb to get the parameters on which the database relies on. In the last section, we will see how to establish such a database with the description of typical recording sessin with Myriam Gourfink..

3.1.2 The database

The database I have been using is actually a set of 40 videos of dance fragments.The hardware configuration to get these videos consists of a DV camera with a 300*200 resolution and at a rate of 25 frames per second.

Figure 3.1: Snapshots of the video A1

35 parameters are extracted for each frame of each video, and then can be loaded in Matlab as a matrix of parameters.
This set of videos consists of 4 different dancers from the Hervé Robbe company. The names of these fragments are as shown below:

dancer 1: A1
dancer 2: A1bis
dancer 3: A2
dancer 4: A2bis
This nomenclature goes from letter A to letter J.

Depending on what we are actually looking for, we can use this database in three different ways (maybe there are some more...) :

-The simplest way is to analyse the 40 videos one by one and not take care of the fact that it is the same choreography or the same dancer than in any other video.

-An other way to analyse data is to establish a comparison between two dancers executing the same choreography, e.g. for example A1 and A1bis. This is a particulary good approach to compare segmentations for example and a good way to check the robustness of a segmentation based on learning-methods.

-The third way to analyse data is to compare the ten different choreographies of the same dancer and try to highlight differences or similarities between dancers.

Each fragment lasts around 30 seconds.
The video camera is still so that there is no changeable influence between the angle of the camera and the data.

To end up with the different ways of considering the database, we can say that we have :
-40 examples of the same choreographer, e.g. the same vision of contemporary dance
-4 by 10 examples of each dancer
-20 by 2 examples of the same dance/choreography

I have an other set of videos that could be analysed via Eyesweb but I didn't yet manage to use it. This set of videos is grabbed from an article which references are listed in the bibliography [DB]. The experiment described in this paper is particularily interesting because it gives us the way the dancers and the choreographer would achieve the segmentation by visualizing their own videos.

3.2 Dealing with Eyesweb

In my situation, Eyesweb provides a list of 35 parameters for each frame. These parameters include:

3.2.1 An estimation of body parts: the pseudo-skeleton

To get these parameters, Eyesweb tries to match a human skeleton to the image silhouette by dividing the blob in multiple areas and by computing the centroid of each area.
Here is the list of the skeleton's parameters:
- Bounding rectangle x,y,width,height
- Head x,y
- Center of gravity x,y
- Left Right Hand x,y
- Left Right elbow x,y
- Left Right shoulder x,y

- Left Right knee x,y
- Left Right foot x,y

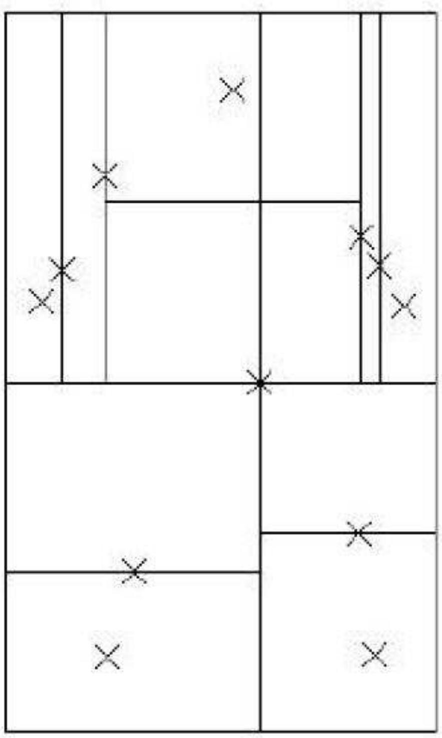


Figure 3.2: Blobs centroids used to compute the skeleton

Most often, the dancer does not face the camera, therefore the computed points hardly match the human body parts. Anyway they are still interesting to be considered as a multi-resolution description of the blob.

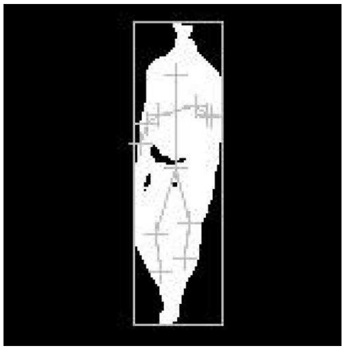


Figure 3.3: Skeleton on a real figure

3.2.2 The mid-level parameters

One of the mid-level parameters available is the **stability index I2S**.
Stability is a number associated to a body posture that attempts to measure its stability (or instability). Roughly, it is computed by dividing the height of the body barycenter by the distance between the feet while they both are on the ground.
Its value can be interpreted in several ways, for example if it is contained in a certain range it indicates a stable position, above or below that range it shows that the position may be unstable.
In this way it is possible to evaluate what kind of positions are used during a performance and how the performer skips through them.
The most direct use, the simplest possible, is detecting peak values of the parameter, corresponding to a step being performed.
Steps frequency is another parameter that can be studied. The picture shows the behaviour of the stability parameter during a dance fragment.
Peak values correspond to feet crossing, making a step.

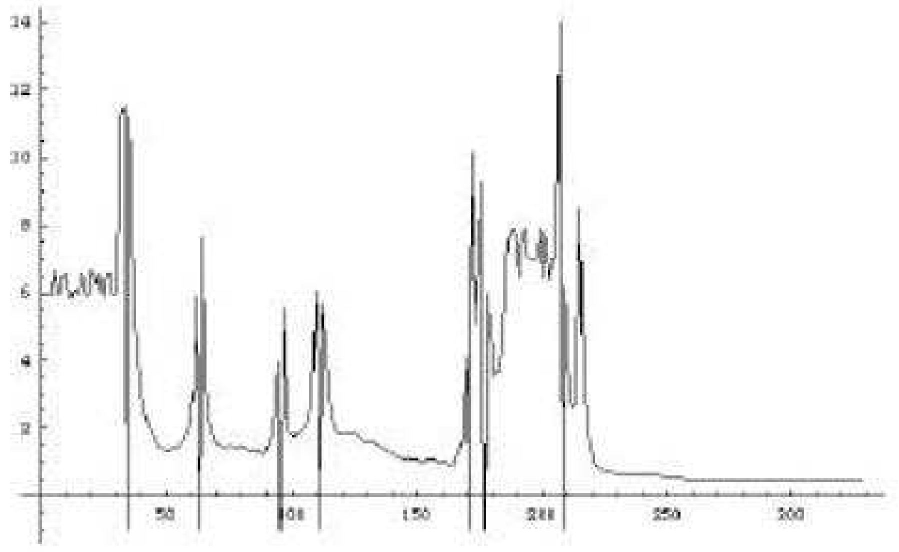


Figure 3.4: Footsteps detection with the stability index

The **expansion or contraction** of the space used by the dancer has been modelized according to the Laban Theory. The Laban's Theory of Effort says the human body is surrounded by a sphere, called Kinesphere, whose amplitude corresponds to the maximum extension of the limbs of the dancer.During a sequence of movements the limbs can extend and touch the outer limit of the Kinesphere, or be kept close to the body.
The **contraction index** is a measure, ranging from 0 to 1, of how the dancer's body uses the space surrounding it. Consequently, it uses a rectangle which can be considered as a bounding box which surrounds the dancer's body. The contraction index is actually a ratio of the area of the silhouette (the number of pixels covered by the silhouette in the frame) and the area covered by this bounding box.

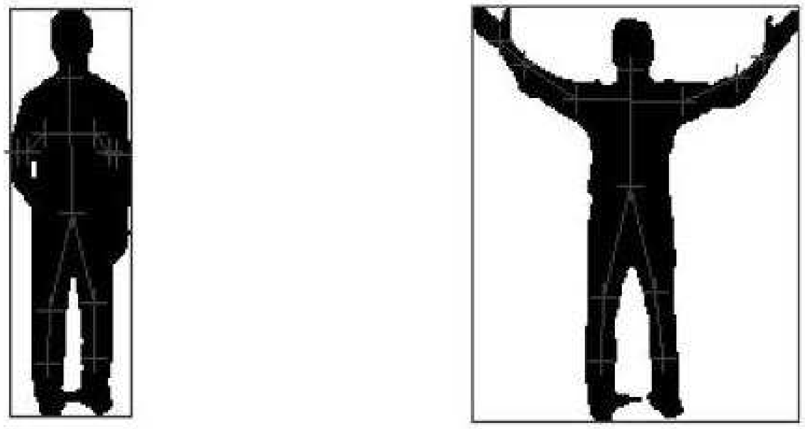


Figure 3.5: Left: index=1 right: index=0

Quantity of Motion (QoM) is computed as the area (i.e., number of pixels) of a Silhouette Motion Image (e.g., the number of pixels in the grey area in the next figure). A Silhouette Motion Image (SMI) is an image carrying information about variations of the silhouette shape and position in the last few frames (see Volpe dissertation[Vol03]).

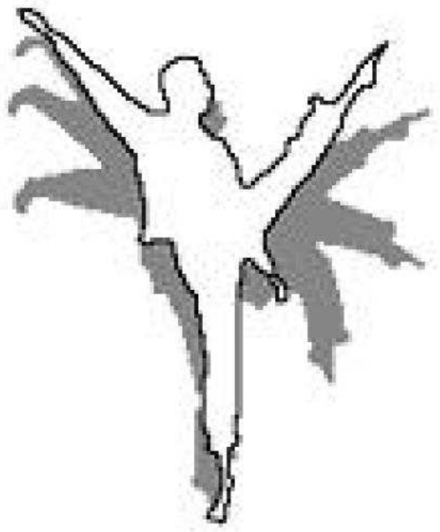


Figure 3.6: A Silhouette Motion Image with time window n=4 frames

Quantity of motion can be considered as an overall measure of the amount of detected motion, involving velocity and force. QoM can be thought as a first and rough approximation of the physical momentum $q = mv$, where m is the mass of the moving body and v stands for its velocity. In the next section we will see how to establish a database to perform non real-time gesture analysis during a test session with Myriam gourfink.

3.2.3 A typical test session with Myriam Gourfink

I have participated to one test session with Myriam Gourfink. Her set of sensors combines optical fibers, breathing and flexion sensors and an accelerometer connected to a wisebox she is wearing at her belt.

The accelerometer (red) is placed on the head.
The breathing sensor (red square) is attached on the chest.
Flexion sensors and optical fibers(blue) are located on ankles, knees, elbows and shoulders to measure the angles of the different body parts as shown by the figure below:

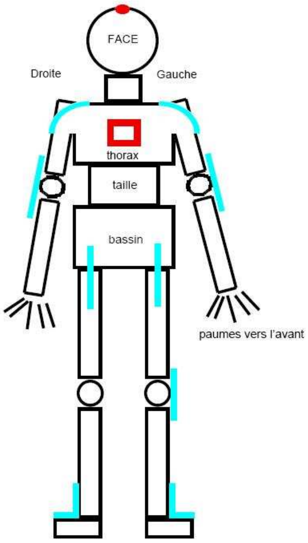


Figure 3.7: The set of sensors used by Myriam Gourfink

The Wisebox is linked to a G4 macintosh via a wifi receiver. The data is then stored with the use of patches made for the real-time environment software MAX/MSP. 3 patches were being used:
-The *Wifi Station* patch aims at calibrate the data transmission.
-The *Sensor Recorder* patch stores the data.
-The *GFKrecognize* patch is based on Hidden Markov Models and does the recognition of movements.

Using this system, several phrases of movement have been recorded according to the conceptions of dance of Myriam Gourfink. She describes simple movements of parts of her body such as arms and legs with different levels of intensity. For example, she can move her left arm in 3 different ways: **mou** (low muscular activity), **medium**, **tonique** (high contraction of the muscles).

A camera has also been used to record the dancer so that we still can use Eyesweb to analyse the video.

Part II

Data analysis and Results

This part of my report deals with most of the work I have been doing at Ircam and highlights some of the results I have been able to find. According to the mainframe of my internship at Ircam, I have tried to program an interface that performs small tasks such as visualizing the data and apply signal-processing techniques on the database.

To my opinion, that way of programming has two main advantages:

-This interface is reusable and could serve as a user-friendly gesture-toolbox in which it could be possible to add newly developped processing techniques. Moreover, with several improvments it could fit to other databases obtained with Eyesweb and videofiles. I think for example of the set of videos given by the choreographer Scott De Lahunta whose analysis could be very interesting according to the fact that we have manual segmentation informations of the choreographer and his dancers provided with the article.

-That way of computing enables the programmer to use and reuse small functions of all the processing techniques available and establish quite easily other strategies of segmenting the parameters by implementing functions of functions.
Consequently, one of the first things I did is to visualize the different parameters provided by Eyesweb.

Chapter 4

Data Visualization

We mentionned before the list of the different parameters provided by Eyesweb.
Here are the different visualizations of the parameters :
The visualization of the skeleton is necessary because it shows via an animation how well the movement of the dancer is capted by Eyesweb. This visualization is based on the animation of the twelve points in a x-y plot which refreshes at a rate chosen by the user and is ideal to see the skeleton's points and trajectories.
The default rate is the video one: 25 frames per second.

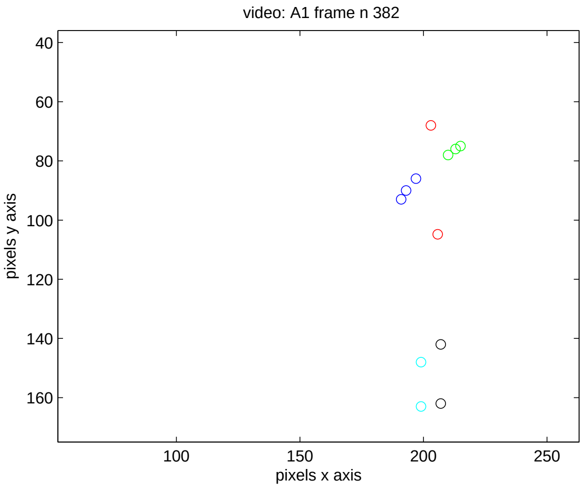


Figure 4.1: The skeleton'sparameters visualization

We can also observe the trajectories pursued by the center of gravity:
We can visualize it as a simple parameter or animated, it is up to the user.

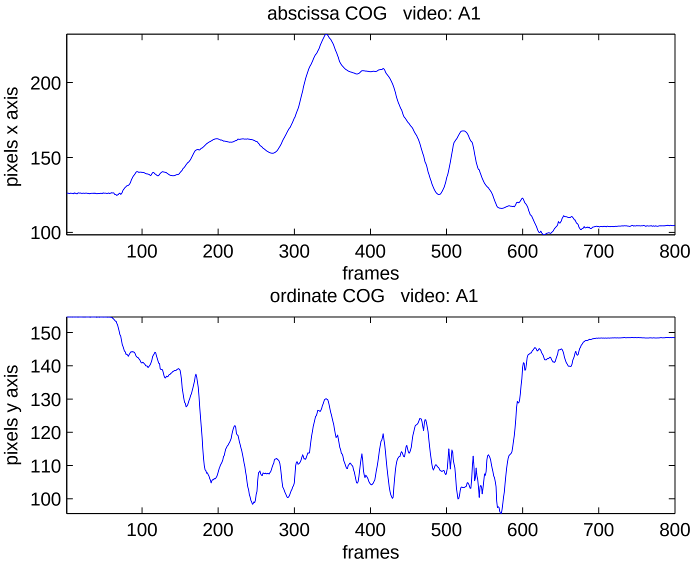


Figure 4.2: The coordinates of the center of gravity

For example, the animated visualization is particularily adaped when comparing two dancers executing the same choreography. This visualization is quite efficient to know if the dancers (their center of gravity) are following the same trajectories or not. The blue point is the trajectory of A1 and the red one of A1bis. We can also observe the differences in the trajectories of the two dancers by supploting the evolution in time of the coordinates of the center of gravity shown below:

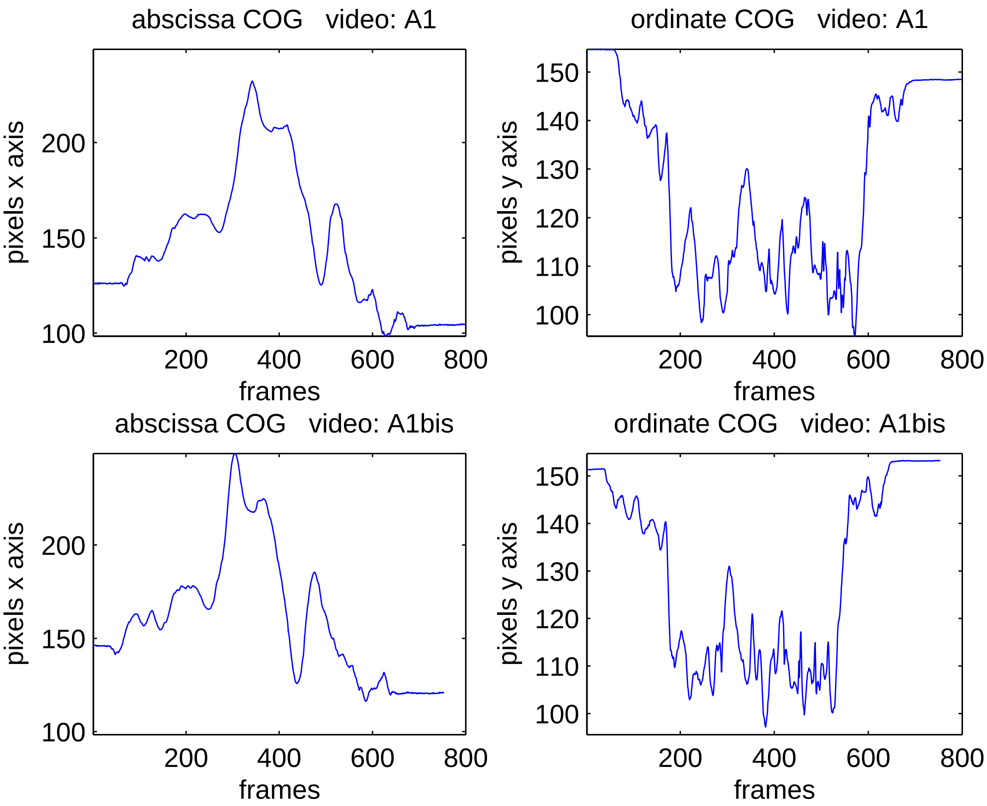


Figure 4.3: The center of gravity: A1 and A1bis

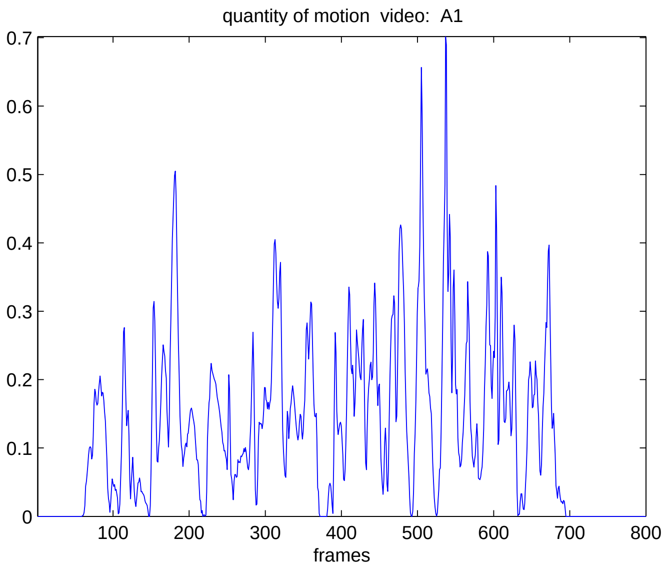


Figure 4.4: The quantity of movement

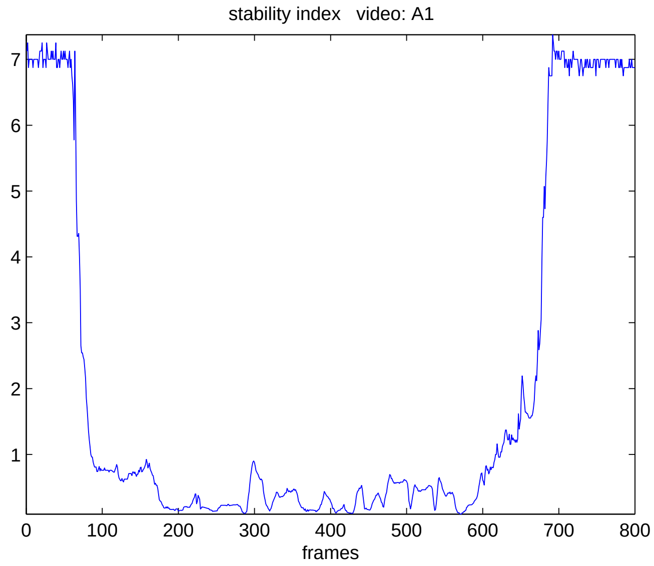


Figure 4.5: The stability index

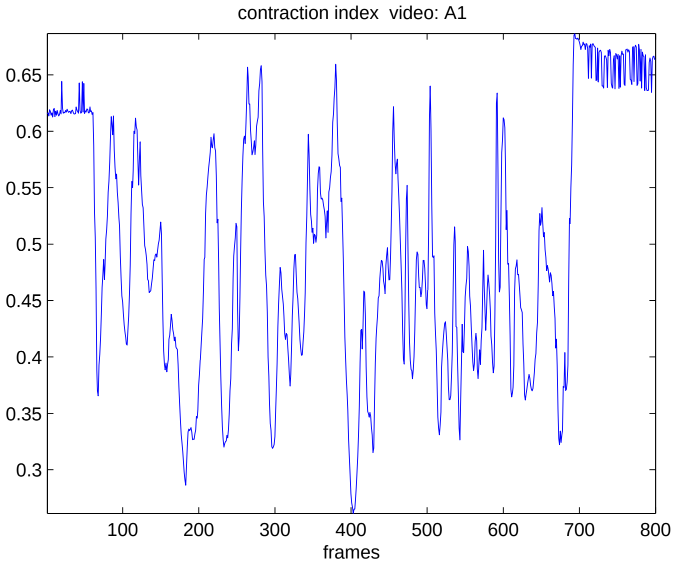


Figure 4.6: The contraction/expansion index

Chapter 5

Principal Components Analysis

One of the first tools I had to deal with during my internship is Principal Components Analysis. Principal Components Analysis is a commonly used statistical technique whose major effect is to reduce and decorrelate the data by taking its redundancy off. This way we can try to identify new meaningful underlying variables by rearranging the data and discover its real dimensionnality.

5.1 Principle

Let us define a space that we will call the parameter space. This parameter space is defined by a set of parameters $\{\mathbf{P}_N\}$ and can be represented by the matrix P with N the number of parameters.

$$P = [\mathbf{P}_1\mathbf{P}_2 \dots \mathbf{P}_N] \tag{5.1}$$

To achieve PCA on this set of parameters, we have to consider its N by N covariance matrix in mean-deviation form.

$$P_{cov} = P'P'^T \tag{5.2}$$

with P' such as the data is centered

$$P' = P - P_{mean} \tag{5.3}$$

Applying PCA to P_{cov} is in fact very similar to eigenvalue decomposition:
We diagonalize P_{cov}

$$Q_{cov} = S^T P_{cov} S \tag{5.4}$$

so that the new covariance matrix Q_{cov} is orthogonal and uncorrelated in the component-space (the eigenvalues of Q_{cov} are disposed on the diagonal and are in order of decreasing variance).

$$\mathbf{Q_{cov}} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_N \end{pmatrix}.$$

To obtain the matrix S(which allows us to establish the change of space and by this way the diagonalization process) we look for the eigenvectors of P_{cov} by calculating the eigenvalues λ of P_{cov}. Each column of the matrix S represents one unit eigenvector. Each parameter P_k is transformed into a principal component Q_k by the orthogonal transformation: Q_k is the coordinate vector of P_k in the component space with respect to the columns of P. Because of the influence of the first eigenvalues (they have the biggest variance), the best way to reduce data is to keep only the first components.

$$Q_k = S^T P_k \tag{5.5}$$

5.2 Application to the database

5.2.1 Reduction of the amount of data

According to the description of my database, each video is described with a set of 35 parameters. If we pay attention on the 24 skeleton's parameters which are an estimation of the body silhouette of the dancer for each frame of the video, we can apply PCA to discover the main trajectories of the skeleton in a dance fragment.

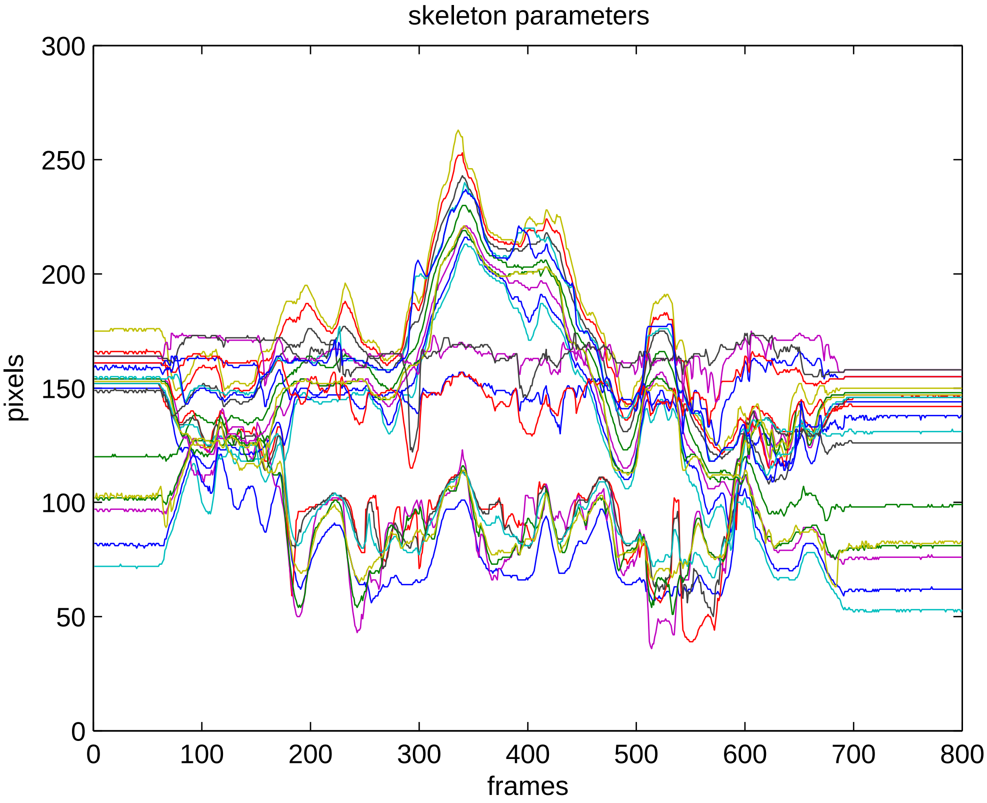


Figure 5.1: The 12 2D-coordinates of the skeleton for the video A1

The first main application of PCA is the reduction of the amount of data. The way I programmed my interface allows the user to choose, according to their importance, how many components he wants to keep in the component space. Applying PCA to the skeleton's parameters, the dimension of the parameter space is 24 and can be reduced to 8 or 9 components. Indeed I use the function *PCAcov* in Matlab which gives a percentage of the most influential components. By this way we can easily choose how many components we want to keep.

The 12 2D-coordinates of the skeleton shown on the figure above are estimation of body parts of the dancer and are therefore very correlated to the center of gravity of the dancer. Consequently, if we look at the first components found by PCA they will look like very strongly to the coordinates of the center of gravity.

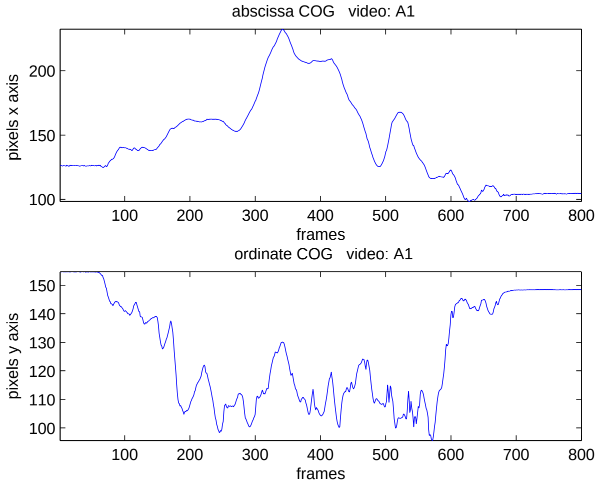


Figure 5.2: The coordinates of the center of gravity

Below is the graph of the two first components. We can notice that they are very close but opposed in sign.

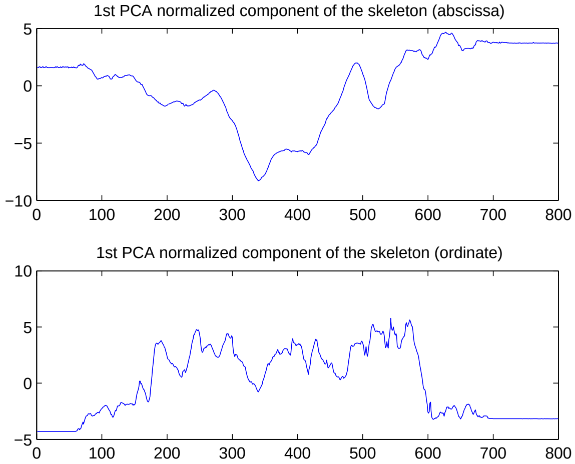


Figure 5.3: The two first principal components

Now we have the same components but opposed. They seem to be almost the same as the coordinates of the center of gravity.

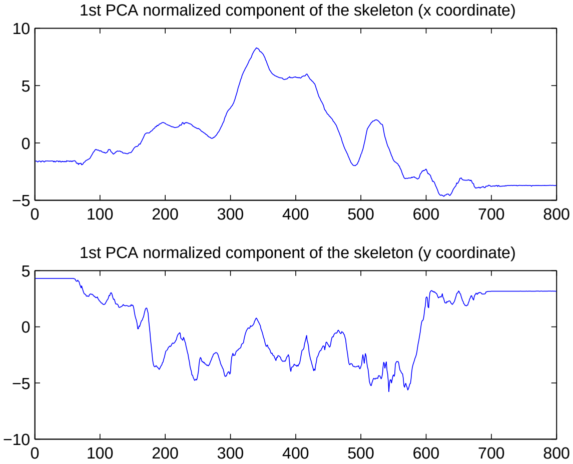


Figure 5.4: The same components but opposed

That is why it could be interesting to substract the coordinates of the center of gravity before applying PCA to see what is actually hidding behind because it is quite useless to perform PCA and establish a segmentation if we can have the same result with the coordinates of the center of gravity without performing PCA.

5.2.2 The different choices offered to the user

In my interface, several choices are proposed to the user including the one of substracting the coordinates of the center of gravity. Indeed by performing this substraction, we can see the movements of the body without having the influence of the trajectories of the center of gravity. In this manner, we can highlight movements that are not correlated to the center of gravity or at least that were not clearly visible.

An another choice that the user can do is to standardize every parameter by its standard deviation. Performing a principal component analysis on a standardized data matrix has the same effect as performing the analysis on the correlation matrix (the covariance matrix from standardized data is equal to the correlation matrix of these data).

Until now, we focused our work on applying PCA on all the parameters of the skeleton indinstinctely. What could be also interesting should to apply PCA on different parts of the skeleton according to morphological considerations of the human body. Consequently, the interface can apply PCA on different parts of the human body such as:

- the arms are defined by the 2D-points corresponding to the shoulder,the elbow and the hand
- the legs are defined by the foot and the knee
- the center of gravity and the head

Symetric or parallel approaches have also been performed such as :
- both arms, both legs
- right arm-left leg or left arm-right leg

Using these last possibilities we can compare the components obtained for the 2 arms of a dancer.

An application of this framework would be to use in a more efficient manner PCA. In Rémy Muller's dissertation [?], it has been shown that we can reduce the skeleton's parameters and do not affect the result of a gesture following using Hidden Markov Models from 24 parameters to 8 or 9 components. Knowing that we can use until 8 or 9 components, the goal of this framework can be seen as finding the best representative/efficient components to achieve the following according to morphological considerations.
Finally, this framework has been applied to a choreography to know if the morphological components are comparable for 2 different dancers executing the same choreography.

Chapter 6

Dynamic Time Warping

In this part, I introduce the basics of a widely used method for automatic alignment: Dynamic Time Warping (DTW [Sch04]). In our context, the use of DTW is necessary because it could allow us to perform comparisons of segmentations between the same parameters of videos from the same choreography. For example, we should be able to compare segmentations of the same parameters between two different dancers executing the same choreography. Refer to chapter 8

6.1 Principle

This technique actually finds the best global alignment of two sequences, based on local distances. It uses a Viterbi path finding algorithm that minimizes the global distances between the sequences. Dynamic Time Warping applied to speech recognition is described in detail in (Rabiner and Juang 1993). Alignment by DTW is carried out in the following steps:

6.1.1 Calculation of local distances between the 2 parameters to be aligned

The local distances are stored in the local distance matrix $ldm(m, n)$ where each value expresses the dissimilarity between the frame m of the parameter 1 and the frame n of the parameter 2.
This dissimilarity matrix is computed by multiplying the 2 parameters so that the size of the matrix is $m \times n$.

$$Dissim = P_1^T * P_2 \tag{6.1}$$

By looking at the dissimilarity matrix below, the ideal alignment is located on the top left bottom right diagonal where the measure of dissimilarity is the lowest (in blue).

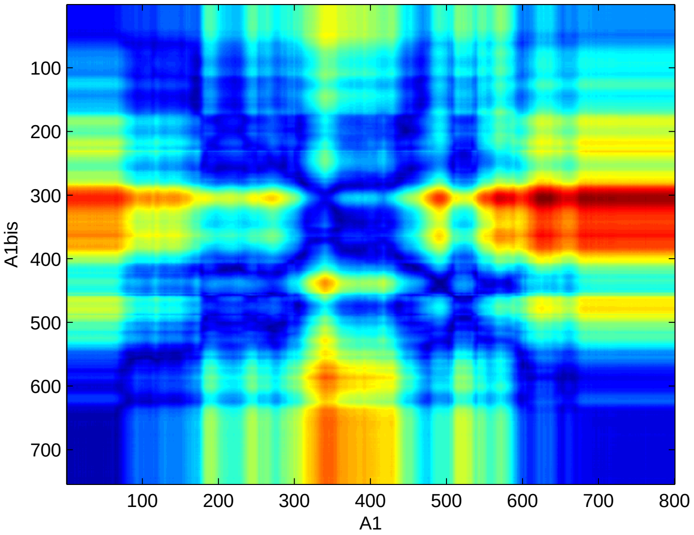


Figure 6.1: The dissimilarity matrix obtained with the skeleton of A1 and A1bis

To compute this matrix, several choices are given to us:
We can use every parameter of the matrix, i.e. 35. On the opposite, we can only take the parameter that we want to be aligned. Indeed it would seem obvious that to get an appropriate alignment of the quantity of motion between $A1$ and $A1bis$ by expressing the dissimilarity matrix with these only two parameters.
It appears that the best choice consists of using the skeleton's parameters without the coordinates of the center of gravity.
Why not moving the center of gravity (cog) ? As a matter of fact, I already said that the skeleton is very correlated to the cog. Consequently, add the cog would be redundant.
With the same idea, it is not necessary to use the whole 35 parameters which are partially redundant. Moreover some of these parameters are so different (I would say non homogenous) that the gain of information would be useless. Therefore, I decided to use the skeleton's parameters because it appears to be quite an homogenous and complete set of parameters to get the dissimilarity matrix. To come to these conclusions, I checked it out by visualizing the data. Indeed the main weakness of DTW is that disturbing effects appear on the data when the dissimilarity between the 2 parameters is huge. Indeed, when the dissimilarity is big, the algorithm is stuck in a valley, i. e. it does not find the optimal path on the diagonal. Therefore, it goes trough the horizontal/vertical way and keeps the same frame m or n (depending on if it goes vertically or horizontally). The result on the parameters is that they

remain constant because they have the value corresponding to the frame m or n. I chose the set of parameters on which this effect had the lowest impact on the data. This disturbing effect could also have been diminished using an other type of local constraints. I have only tested the type one (see next section). Here are some examples of these effects on the data:
If you compare the next figure with the one of the quantity of motion of $A1$ in chapter 6, you will see those effects of DTW.

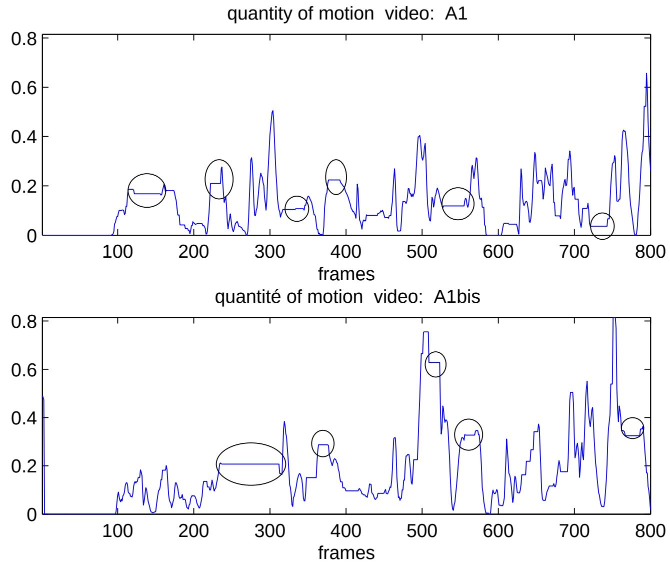
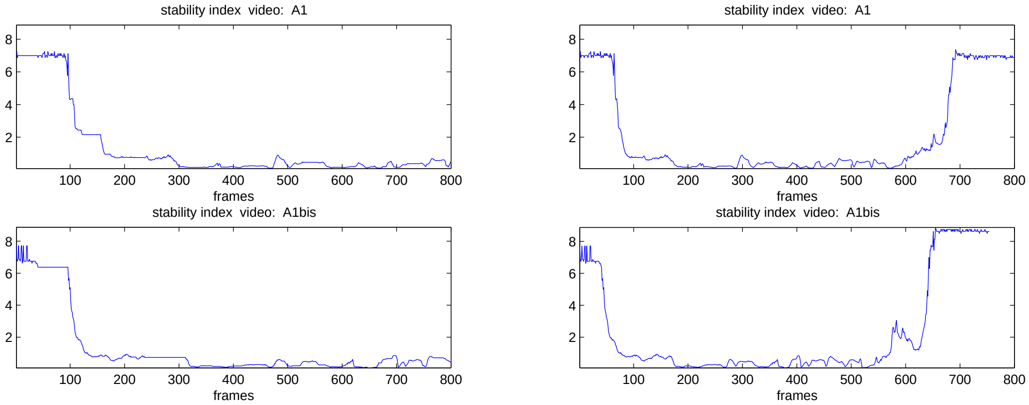


Figure 6.2: The effects of DTW with the 35 parameters used to get the dissimilarity matrix

This is more visible with the stability index which is completely deformed.



(a). The effects of DTW on the stability index with the 35 parameters used to get the dissimilarity matrix

(b). The regular stability index.

Figure 6.3: The effect of DTW established with the 35 parameters used to get the dissimilarity matrix.

6.1.2 Calculation of local and global constraints

This part deals with the calculation of local and global constraints to establish an augmented distance matrix $adm(m,n)$.

This matrix corresponds to the global cost (the accumulated global distance) of the best path up to the point (m,n). The alignment is given in the form of a path through the augmented distance matrix. If a path goes through (m,n), this means that the frame m of the parameter 1 is aligned with the frame n of the parameter 2.

The local path constraints define how the augmented distance matrix adm at point (m,n) is calculated from the local distances, with $ldm(m,n)$ abbreviated to λ.

$$adm(m,n) = \min \left\{ \begin{array}{ll} adm(m-1,n-1) & + w_d\,\lambda \\ adm(m-1,n) & + w_v\,\lambda \\ adm(m,n-1) & + w_h\,\lambda \end{array} \right\}$$

Figure 6.4: The augmented Distance Matrix at point (m,n) with the type 1 local path constraints

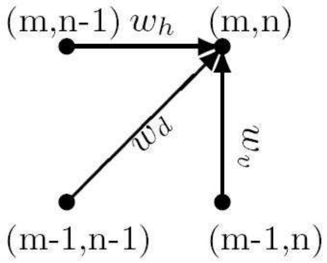


Figure 6.5: Neighbourhood on point (m, n) type 1

The constraint type 1 allows horizontal and vertical steps and thus admits extra or forgotten events.
I chose $(w_h, w_v, w_d) = (1, 1, 1)$ for the weights of the local path constaints.

6.1.3 Finding of the optimal alignment path

This path minimises the global distance by the Viterbi algorithm.
The DTW algorithm finds the best alignment path between the two parameters according to the local distances and the local and global constraints calculated above. The best path is computed iteratively, by updating the augmented distance matrix $adm(m, n)$, which is the cost of the best path up to the point (m, n).

6.2 The effect on the data

Using DTW to align the different parameters, here are some visualizations of the set of parameters aligned for the choreography including $A1$ and $A1bis$.

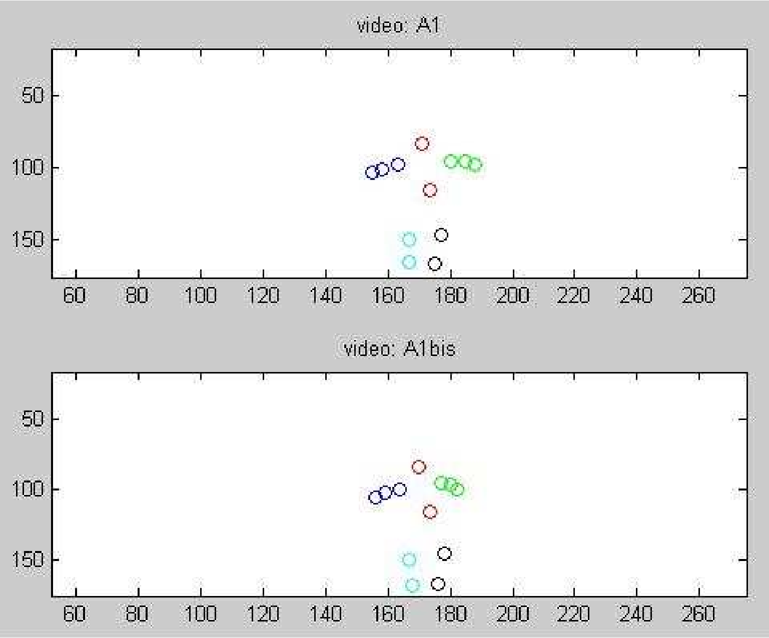
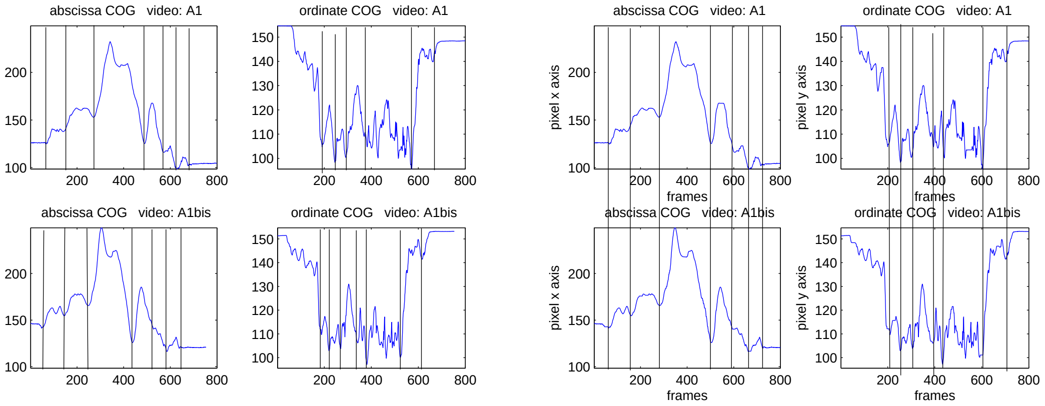


Figure 6.6: The alignment of the skeletons

This picture is not the most appropriate way of checking that the two skeletons are doing the same movements at the same time. Anyway, via the animation mode, the two skeletons have the same gesture , the same position in the screen of the camera and at the same time.
Because of the fact that I calculated the dissimilarity matrix with the skeletons, this is the case where DTW is the most efficient.

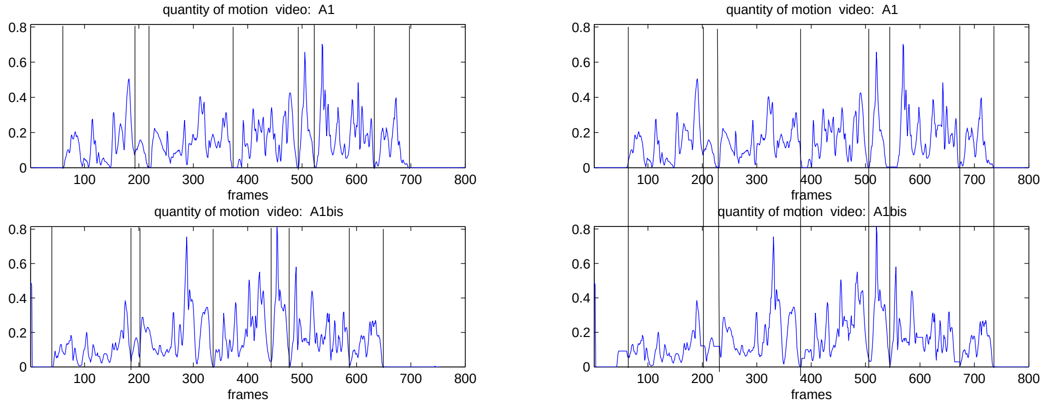
Below is the effect of DTW on the coordinates of the cog. Because of the fact that the cog is correlated to the skeleton, this is the second most efficient use of DTW. I put significant lines on the graphs to highlight the alignment between A1 and A1bis. The way I chose the lines is not heuristic. I placed them at the minimum values of recognizable shapes so that it is possible to evaluate the alignment performed with DTW. The alignment of the other parameters such as the quantity of motion and the contraction index are not as good as the alignment of the parameters above. This is due to the fact that they are less correlated to the skeleton. Anyway, this is still good enough to perform comparisons between segmentations.



(a). The coordinates of the COG.

(b). Alignment of the coordinates of the COG.

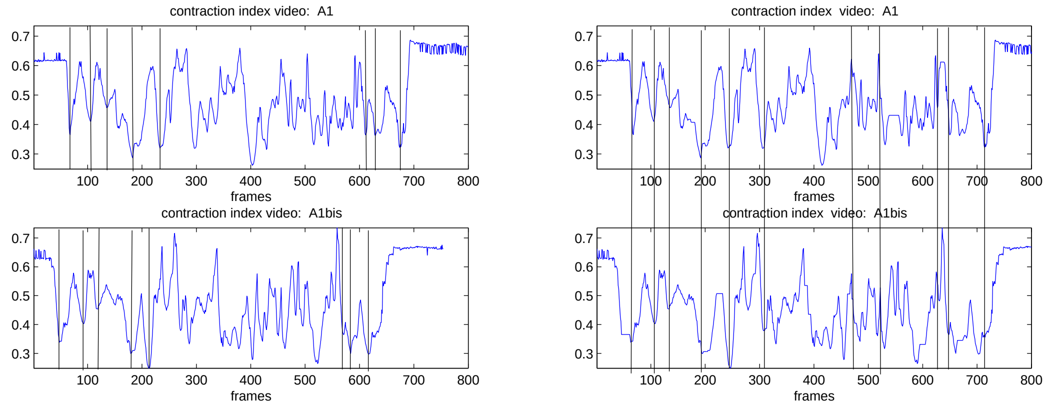
Figure 6.7: Effect of DTW on the COG.



(a). The quantity of motion.

(b). Alignment of the quantity of motion.

Figure 6.8: Effect of DTW on the quantity of motion.



(a). The contraction index.

(b). Alignment of the contraction index.

Figure 6.9: Effect of DTW on the contraction index.

Chapter 7

What and how to segment? the segmentation strategies

7.1 The automatic methods to segment data

What can be seen as the automatic way to achieve the segmentation is the detection of singular values of the parameters, the detection of specific parts of the data by applying a threshold which has to be calculated depending on what we want to segment, recognize, highlight in the data. The value of the threshold can be empiric, or a ratio of the range of values that each parameter can have.
In this part, I expose some of these methods that I use to segment data.

7.1.1 The use of thresholds

Camurri and collaborators (InfoMus Lab Genova Italy) proposed the use thresholds to achieve the segmentation of one of their mid-level parameters which is the quantity of motion.
Indeed they use this parameter as an indicator of movement. They stipulate that if the value of the amount of movement is higher than 25 percent of the maximum value (which is 1 because this parameter is normalized and has its range of values between 0 and 1), there is movement.

7.1.2 The zero-crossing rate

This method calculates for each frame the sign between the value of the parameter and the previous one. If the sign is negative, then it means that the parameter has crossed the zero axis.

7.1.3 The use of minimum/maximum value detection

It exists several ways of determining peak values, singular values of a parameter. The one I suggest here is based on SGOLAY smoothing filters. Savitzky-Golay smoothing filters (also called digital smoothing polynomial filters or least squares smoothing filters) are typically used to "smooth out" a noisy signal. In our case the noisy signal can be PCA components, position of the center of gravity or the different mid-level-parameters.
The principle is quite simple :
Let us call x the input of the filter corresponding to the parameter. x is the output and h is the transcient function of the filter.

$$\text{y=h}\star\text{x} \tag{7.1}$$

Savitzky-Golay smoothing filters perform much better than standard averaging FIR filters, which tend to filter out a significant portion of the signal's high frequency content along with the noise. Although Savitzky-Golay filters are more effective at preserving the pertinent high frequency components of the signal, they are less successful than standard averaging FIR filters at rejecting noise. Savitzky-Golay filters are optimal in the sense that they minimize the least-squares error in fitting a polynomial to each frame of noisy data.
In our application, the main advantage of using SGOLAY filters is that it approximates the parameter by a polynomial function. We can choose the size of the window and the order of the polynomial function to approximate the parameter. Therefore, derivates of the parameter can easily be obtained with the approximation.
Depending on the parameters, the values of the windowsize and the order of the polynomial function have been fixed.
The way these values have been fixed depends on how accurate the approximation is. For example, I chose a window size of 17 frames and a filter order of 6 to get a good approximation of the quantity of motion and a good detection of the minimum/maximum values.

7.2 Different strategies to segment the movement

To evaluate a strategy of segmentation, we have to know first what we want to segment. A segmentation has to be applied to a parameter or a motion descriptor. These parameters can be :

- the skeleton or some parameters derived of the skeleton :

  – the position of the center of gravity.

– the principal components

A general human motion following system can be achieved using the segmentation based on the changes of direction of the curvature trajectories of the center of gravity (see Rémy's Muller dissertation [Mul]).
A more precise gesture following can be done with the same segmentation but with different parameters such as the principal components. By using the components found for different parts of the body (see PCA chapter), we can apply the segmentation based on the changes of direction of the curvature trajectories. We could also use the norm of the gradient. But this still needs to be tested.

– the vectors and angles corresponding to morphological parts of the human body.
With this interpretation of the data, I match vectors which have a length of a certain number of pixels with real parts of the human body. By applying this method I have a model of every part of the dancer's body but this model does not represent the reality at all. The length of the vectors considered as body limbs are to close to the pixel size to be considered as significant data.

– the dimensions of the rectangle surrounding the skeleton.

- mid-level parameters :

– the quantity of motion

– the contraction/expansion index

– the stability index

Different strategies can be established to perform segmentation of the movement. Because of the fact that we have no definition of a unit gesture in the context of contemporary dance (unlike the classical ballet), these strategies are established depending on what we want to segment in the data. For example, people from Eyesweb use a method that finds automatically maximum values of the stability index. These maximums are interpreted as footsteps. Therefore, a segmentation can be accomplished to detect footsteps in a dance fragment.

7.3 Static gesture recognition

An other strategy to segment the movement consists of the recognition of static gestures or postures. To be able to achieve this recognition, we first have to look at the different parameters available and try to highlight a typical configuration of these parameters corresponding to a particular posture.

7.3.1 Laid position

If we take the example of the posture which corresponds to a dancer laid on the stage, we have to look in the data if it corresponds to any singular value of the parameters. Once again, there are several ways to do this :

- We can use the skeleton and try to characterize a geometrical configuration of the points.

- We can also use the parameters corresponding to the second and third column of the matrix which are the width and height of the rectangle that surrounds the body. By looking at these parameters during this particular posture, we can observe that the width is at one of its maximum values and the height one of its minimum values as you can see on the figure below.

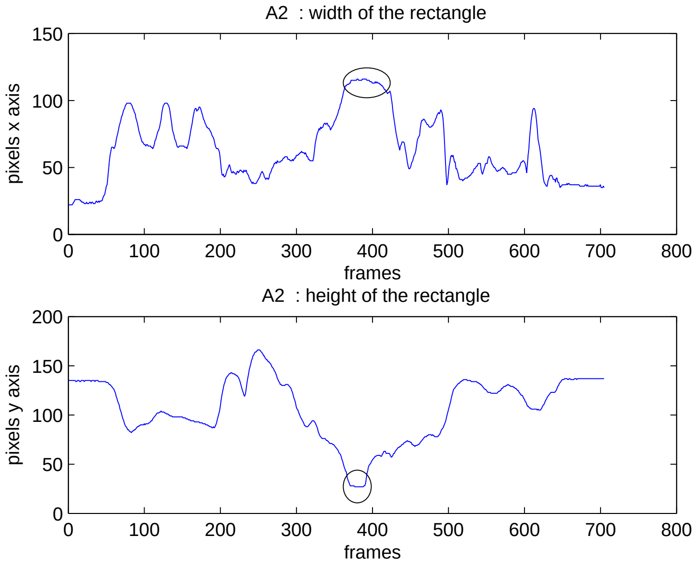


Figure 7.1: The singular values of the rectangle parameter for a laid position of the dancer

- If the dancer is not moving and is laying on the stage, that means that the movement indicator which is the quantity of motion has to be null or quite close to zero. A typical method to do this recognition is the zero-crossing rate. A threshold can also be used but that kind of method is quite dangerous according to the fact that the quantity of motion depends on the distance between the dancer and the camera.

- By looking at the stability index and according to the way it has been computed (see chapter 4), a laid position of the dancer on the stage will match with a maximum value of the stability index.

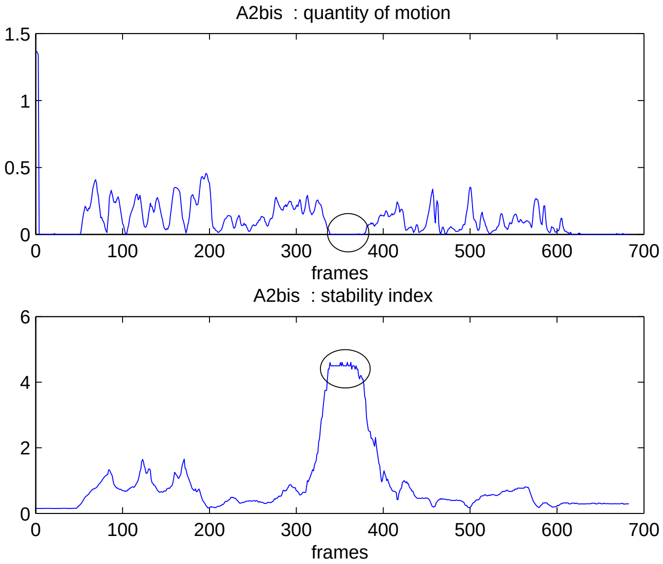


Figure 7.2: The singular values of the Q2M and I2S for a laid position of the dancer

7.3.2 Upright position

In a similar way, we can look for typical values of some parameters to characterize an upright position of the dancer.
- the quantity of motion is close to zero.
- We observe a maximum value for the height of the rectangle and a minimum value for its width.

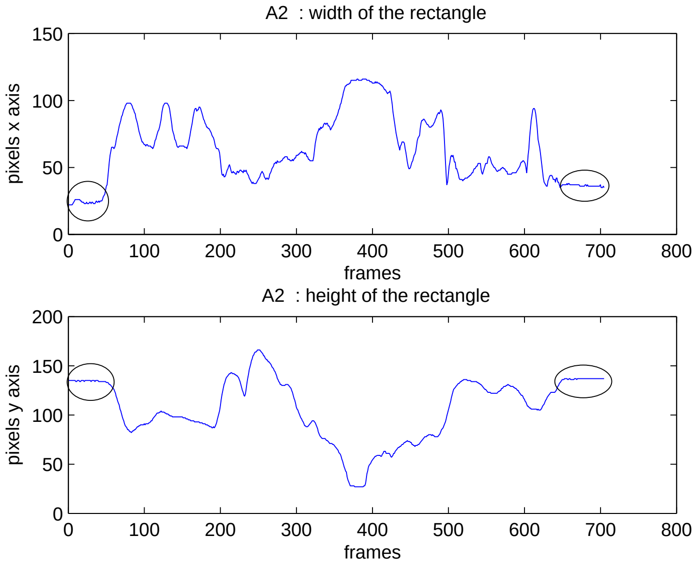


Figure 7.3: The singular values of the rectangle parameter for an upright position of the dancer

7.4 Footsteps detection

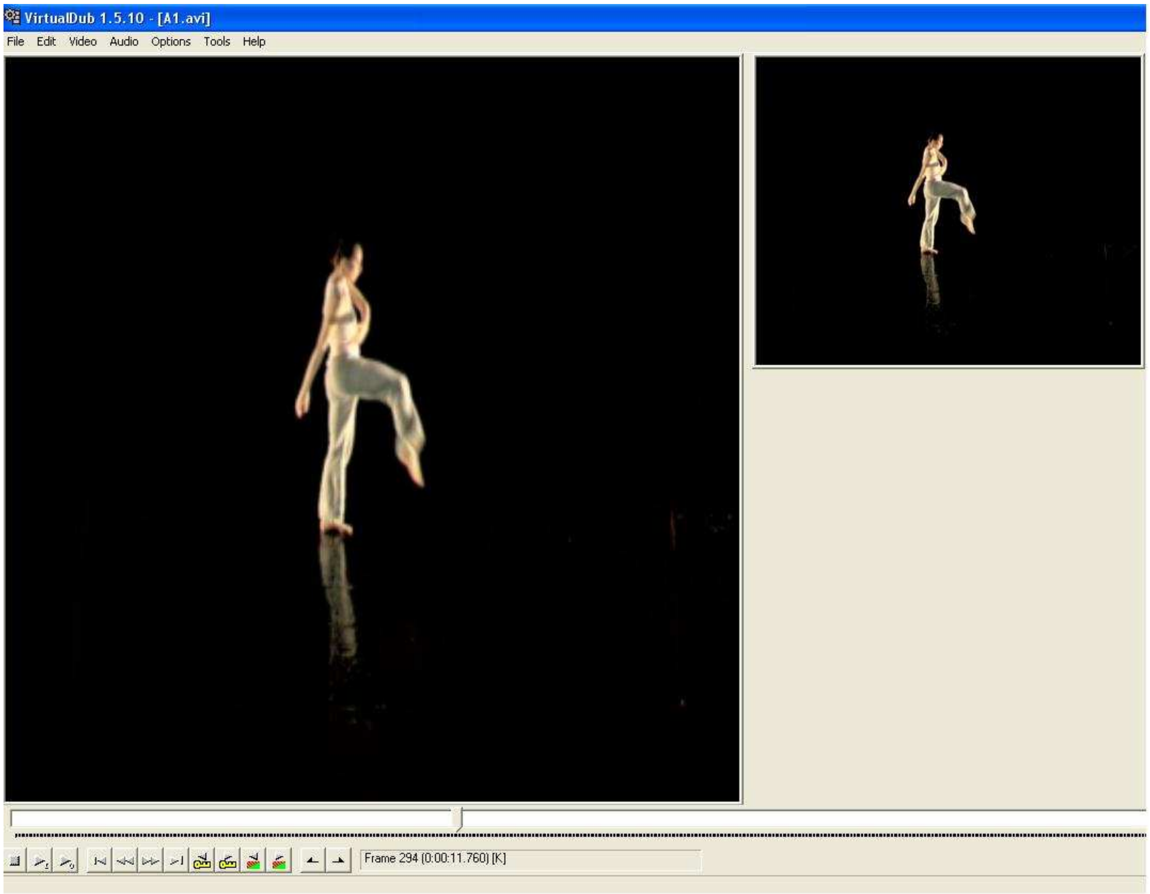
According to an article related to Eyesweb, footstep detection can be achieved by looking at the peak values of the stability index. In my database, such visible peak values corresponding to footsteps do not exist. The peak values I get with sgolay have nothing to do with footsteps. But some minimum values of the quantity of motion can be interpreted as footsteps. Indeed the quantity of motion seems to be a parameter which is representative of a lot of events. If you look next section at the figure where I performed gesture labeling on the quantity of motion of the video $A2$, many interpretations are possible.

7.5 How to measure the goodness/efficiency of a segmentation?

The best way I found to measure how well does a segmentation work was to visualize data.
This has only been possible with the use of appropriate visualization tools.

7.5.1 With VirtualDub

The main advantage of using VirtualDub is that we can look at the videos of the dancers frame by frame. Therefore knowing how the different parameters are computed, we can more easily understand, interpret their singular values by watching precisely at the frames around each singular value.



Frame 294 (0:00:11.760) [K]

Figure 7.4: The Visualization frame by frame with Virtual Dub

By using VirtualDub, I manually segmented the quantity of motion and tried to put significant interpretation to any singular value.
Here is the figure of the interpretation I made of the different singular values by watching at the videos with Virtual Dub.

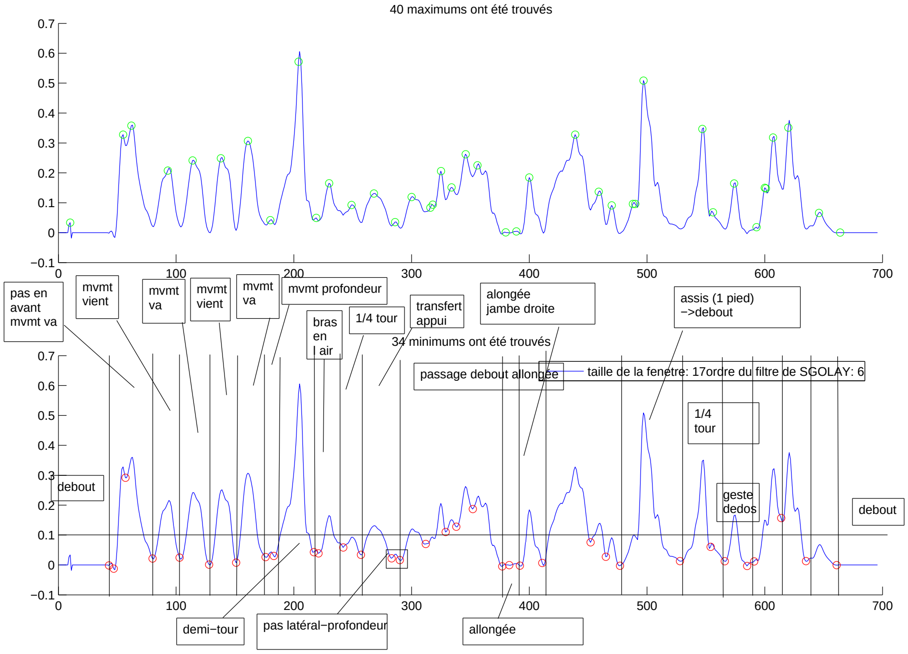


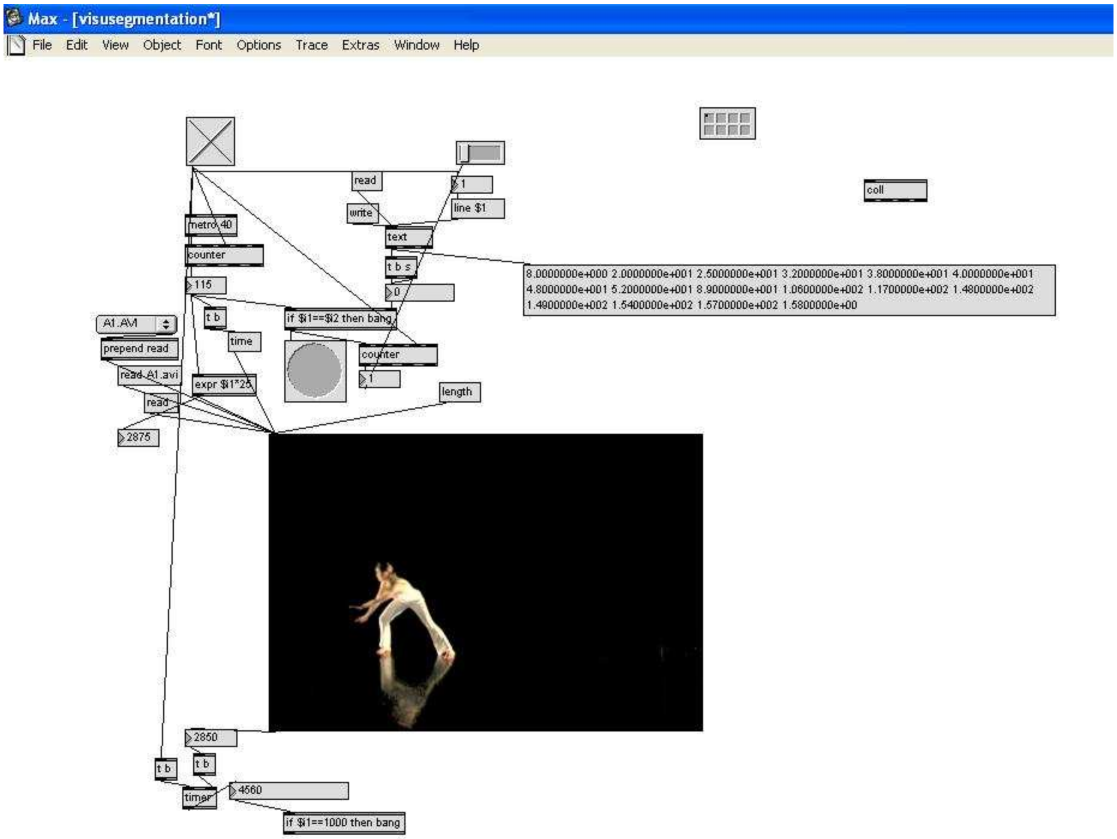
Figure 7.5: The gesture labeling of the quantity of motion

The main interest to perform such interpretation of a segmentation based on singular values such as minimums is that it makes you think of different ways to improve your automatic segmentation. For example, with this gesture labeling of the quantity of motion, I would not use a threshold such as : there is movement if the QoM is more than .25, there is not under. I would try to define different kinds or classes of minimum values with the use of some adaptative threshold. With this kind of methodology, I would differenciate several levels of singular values of the parameter. Indeed, I observe minimums that are relative to changes of direction of the movement (only the movement which is in the field of view of the camera and not in the depth because if you look at some events such as sideways footsteps in the depth of the field of view of the camera -see for example around frame 290, they are not capted with the same intensity as the same sideways footstep in the field of view of the camera) : these minimums would be the ones close to zero(I would mention the activity at the beginning from the frame 50 to 175). To my opinion, I would see minimums that could be interpreted as a body part touching the ground. By looking at the segment between the frames 290 and 375, this segment can be interpreted as the transition where she goes from a still position to a laid down position. To go from one to the other, she puts her left hand, then her bottom and finally her back on the ground. Every part of her body she puts on the ground matches to a minimum value that we can see on the data. This can be explained by the fact that each part of her body she puts on the ground is fixed and therefore makes the indicator of motion diminish. To recognize such transition phases, we should first be able to recognize postures (see previous sections) and we should look at the frequency or the number of singular values between to postures.

7.5.2 With a MAX Patch

The use of MAX has become obvious due to the fact that Matlab is quite in trouble when it deals with video data. The amount of data seems to be too huge to be supported by Matlab in real-time.Therefore it is quite impossible to achieve an interface which allows the user to watch a video frame by frame. To that purpose, we built a patch with MAX which is especially adapted to that kind of applications.

The principle is quite simple. With Matlab, I convert the segmentation list into a a text file in such a way that it is readable by the Max object coll. Then, the patch has been built so that it compares the time elapsed on the video with the list of singular points of the segmentation. Each time these two values are the same, a bang (a flashing light) appears on the screen. A slider has been added to this patch to allow the watching frame by frame.



```
8.0000000e+000 2.0000000e+001 2.5000000e+001 3.2000000e+001 3.8000000e+001 4.0000000e+001
4.8000000e+001 5.2000000e+001 8.9000000e+001 1.0600000e+002 1.1700000e+002 1.4800000e+002
1.4900000e+002 1.5400000e+002 1.5700000e+002 1.5800000e+00
```

Figure 7.6: The visualization frame by frame with MAX

Chapter 8

How to compare segmentations?

8.1 What for?

There are two main reasons to compare segmentations. The first is to know if some events appear at the same time but on different parameters of the same video. The second one is to check if the automatic methods used to perform the segmentation are robust enough to get similar results with a choreography danced by two different dancers. Consequently the comparisons have been established in two main contexts:
- Comparison of the segmentations based on the different parameters of the same video
- Comparison of the segmentations based on the same parameters for the same choreography
This whole framework should be applied to the different strategies of segmenting the data that we have seen in the previous chapter. Until now, it has only be achieved on PCA components and the mid-level parameters.

8.2 Comparison of the segmentations based on different parameters of the same video

8.2.1 Presentation of the Criteria

I actually found three different ways to compare segmentations. These ways consist of calculating descriptors that we would call criteria.
Criteria 1 and 2 are almost identical. They try to describe the correspondance between two segmentations.
A typical example of how the criteria can be used is to compare the segmentations based on the minimums of the Qantity of Motion and of the

Contraction index. A is the list of the minimums obtained with Sgolay applied to the QoM anb B with Sgolay applied to the Contraction Index.

$$criterion1 = \frac{\Sigma_{i,j} min|A_i - B_j|}{length(A)} \tag{8.1}$$

$$criterion2 = \frac{\Sigma_{i,j} min|B_i - A_j|}{length(B)} \tag{8.2}$$

A third criteria has been created to highlight non-detections and wrong alarms. It fills two empty vectors which have the size corresponding to the number of frames of the video. Each vector is filled with zeros and ones according to the two lists of markers that we want to compare: the vector is filled with a 1 at the position of a marker and with zeros elsewhere.

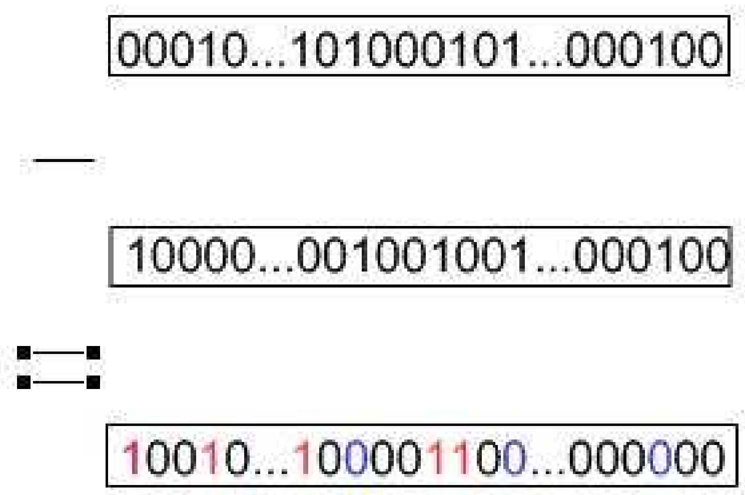


Figure 8.1: The third criterion

The third vector is the substraction of the two others without taking care of the sign. A n frame-tolerance can be implemented by applying the convolution of a normalized n frame-window to the vectors. The addition of ones still left in the substraction vector corresponds to markers that do not have their equivalent in the other vector. The criterion is computed as the ratio of these ones with the number of markers in the two vectors. For a better interpretation, the criterion is modified so that it gives percent success results of how these segmentations match.

8.2.2 Results

Depending on the tolerance/window size that we use, results can be improved. In fact, most of the results I get with the comparison of PCA based segmentation and mid-level parameter segmentation are quite bad. This can be explained by the establishment of criteria to check if 2 segmentations match together. This seems to be a very difficult indicator to perform a good interpretation of the matching between the two parameters.
The comparison of the segmentations based on the mid-level parameters has been more efficient and has allowed some more interpretations of the results. Some work still needs to be done to get some appropriate interpretations of the results but see the annex 1.

8.3 Comparison of the segmentations based on the same parameters for the same choreography

The use of Dynamic Time Warping is necessary because it allows us to compare automatic segmentations based on the same parameter (for example the quantity of motion, the contraction index or the stability index) but for 2 different dancers who are dancing the same choreography. Indeed if we compare the min-max segmentation based on the amount of movement of the video A1 with A1bis, the comparison will be unefficient due to the fact that the videos A1 and A1bis do not have the same number of frames. With the use of DTW, we have the same temporal base for the parameters of the 2 videos. We are now able to compare segmentations obtained with these parameters.

In this context, I propose a methodology which will create tolerance classes corresponding to the distance between two matching points (respectively one-frame-distance, two-frame-distance etc.) In this example, I deal with the minimums of the QoM and I have defined 4 classes.
- tolerance 0: these points match together perfectly.
- tolerance 1: these points have a distance of 1 frame.
- tolerance 2: these points have a distance of 2 frames.
- tolerance 3: these points have a distance of 3 frames.
- the last class includes all the points that have a distance bigger than 3 frames.
A new criterion can be then established depending on how many classes we consider. It is calculated in the following way:

$$criterion = \frac{\Sigma_i n_i}{N} \tag{8.3}$$

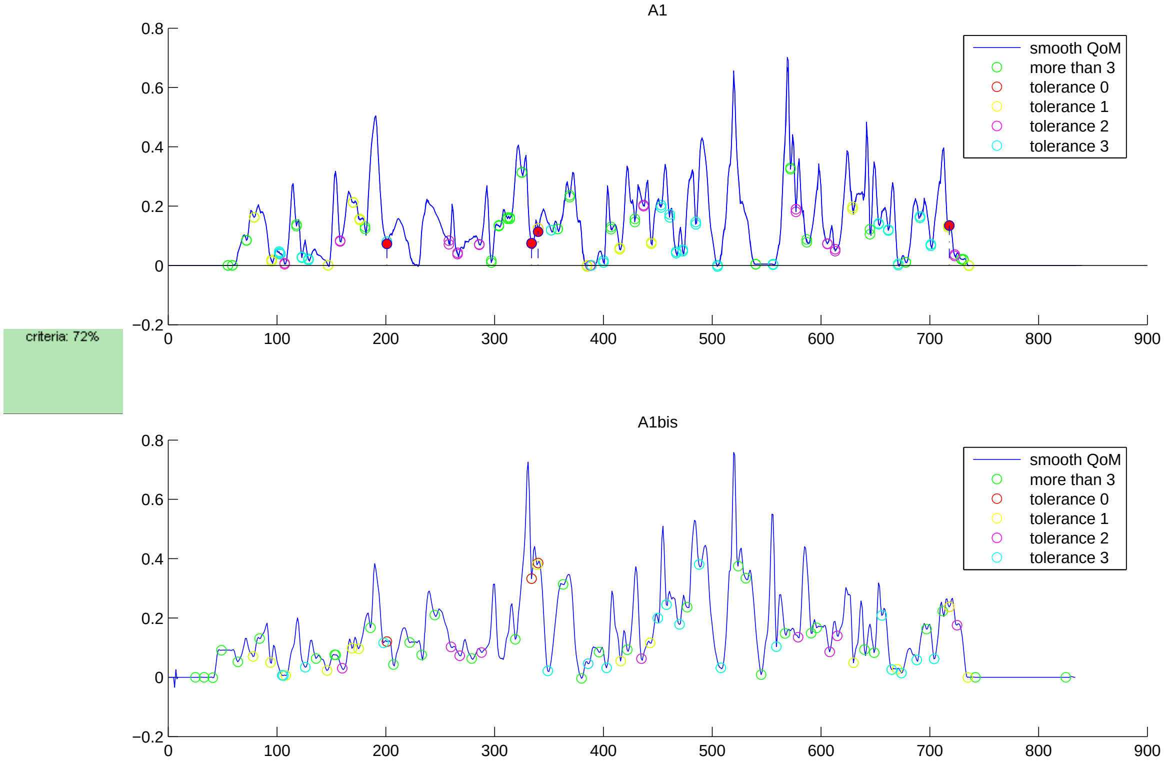


Figure 8.2: Comparison of a minimum based segmentation of the QoM for the choregraphy A1

with n_i the number of points in the class i and N the number of minimums found with the SGOLAY processing step.
I applied this methodology to the Quantity of Motion of every choregraphy of the database.
See the example of $A1$ below (you will find details in the annex 2):

This criterion has an average value of 74.8118 percent. Its lowest value is around 55 percent ant its biggest around 85 percent.

A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
71	73	83	82	74	65	84	81	56	62
F1	F2	G1	G2	H1	H2	I1	I2	J1	J2
70	76	76	77	66	79	78	83	73	82

We have to be very carefull with the use of such a criterion.
First of all, some minimums can belong to two or even more classes at the same time. An improvement of this criterion is to check that no points are considered twice (or more) in this methodology. That modification would decrease the results of the criterion.
Then, the use of DTW to align the parameters modifies the general shape of the parameters so that minimums generated because of DTW are detected. Because of the fact that all the minimums detected by SGOLAY cannot be easily interpreted in regard of the dancer's motion, we are comparing useless minimums that do not deal in the data with the real motion of the dancer.
Finally the results of this methodology is very dependant on the smoothing of the data and thus of the size of the window. In this case, I chose a windowsize of 11 frames.
In general, the larger the size of the window is, the worse the results of the criterion are. This seems to be absolutely obvious according to the fact that with a large window, the smoothing of the data is so strong that it deforms the data.Therefore a segmentation based on these deformed parameters is wrong by definition and cannot represent the movements correctly.
A good application is to look at the choreography A2 on which the gesture labeling has been performed (see chapter 7) and see if the points corresponding to the interpretation of movements match well or not. This labeling has been achieved with a windowsize of 17 frames. The lines drawn on the next figure represent the boundaries of the gestures that have been labeled previously. These boundaries correspond to the points that match quite well (most of them belong to classes 0, 1, 2 and 3).

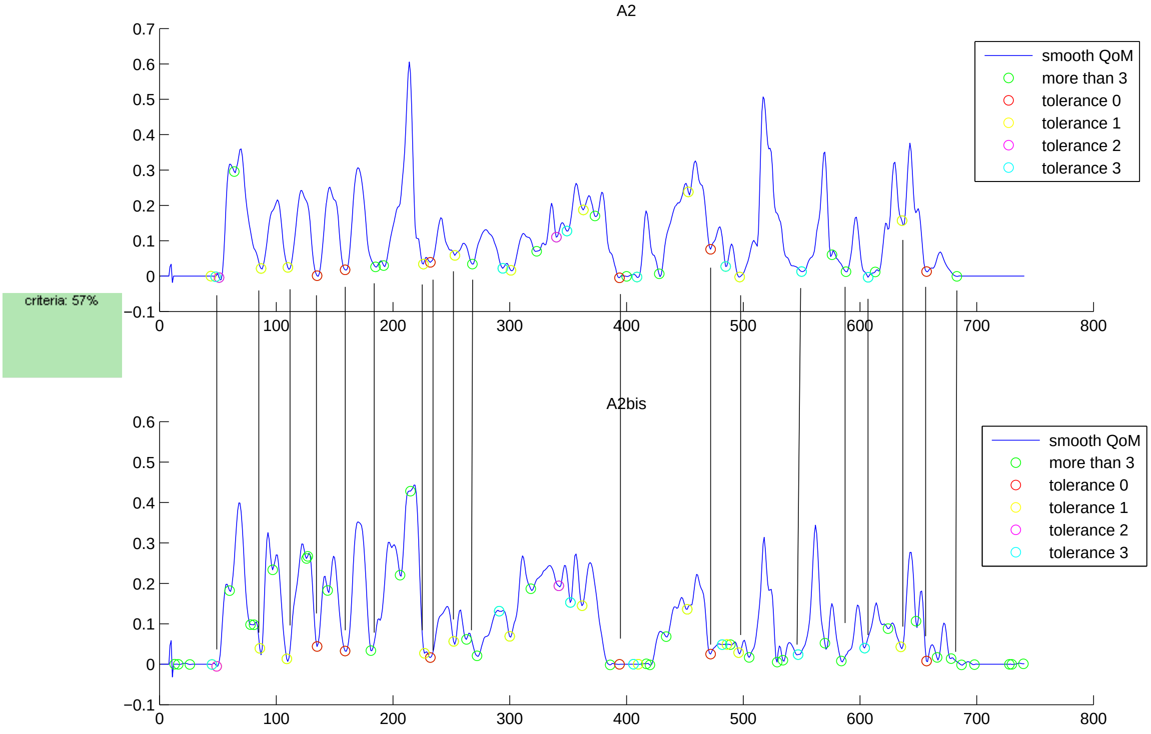


Figure 8.3: Comparison of a minimum based segmentation of the QoM for the choregraphy A2

Conclusion

During my internship at Ircam, I had to deal with both segmentation and decomposition of the movement using computer vision techniques in the context of contemporary dances. To that purpose, I have been working on a database which was in fact a set of 40 videos, each of them including 35 parameters. This database has been established with the use of Eyesweb, a computer-vision based software, and then analysed with Matlab. I have studied the decomposition aspect with the use of a commonly used statistical technique such as Principal Component Analysis. I have tried to perform it in various ways in order to find its most appropriate use according to the robustness of the data provided by Eyesweb.

The segmentation aspect has only partially be studied. I have proposed different methods to automatically segment the data according to an efficient interpretation and use of the different parameters availables. These methods only refer to an automatic way to achieve the segmentation. I have tried to measure how well these segmentations could fit to different kinds of recognition process. Finally I have established comparisons between the segmentations of the parameters from a single video first, and then with the use of a tool such as Dynamic Time Warping, between the same parameters of 2 different videos corresponding to a given choreography.

At last but not least, the whole framework listed above has been done generically so that it is possible to do these tasks with any video of the database using Matlab. Moreover, this way of programming has allowed to give some general results, calculated on the whole database. This has been particularily usefull for example to get general results from the criteria for example (a statistical analysis can be performed to know if the criteria are good or bad estimators of the efficiency of a segmentation).

To conclude with what has not been done yet, I would like for the time being left to get some results from all the segmentations I have listed and be able to establish comparisons to know what segmentation is the best for every parameter. Consequently, the use of a learning method such as belief networks could then be done to combine all the different approaches of segmenting the data to achieve a system capable of segmenting the movement depending on the choreographical context.

Bibliography

[ACLS94] J. Aggarwal, Q. Cai, W. Liao, and B. Sabata, *Articulated and elastic non-rigid motion: A review*, 1994.

[BD02] Gary R. Bradski and James W. Davis, *Motion segmentation and pose recognition with motion history gradients*, Machine Vision and Applications, vol. 13, Springer – Verlag, 2002, pp. 174 – 184.

[Bob] Aaron F. Bobick, *Movement, activity, and action: The role of knowledge in the perception of motion (1997).*

[BP02] Ari Y. Benbasat and Joseph A. Paradiso, *An inertial measurement framework for gesture recognition and applications*, GW '01: Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (London, UK), Springer-Verlag, 2002, pp. 9–20.

[DB] Scott DeLahunta and Philip Barnard, *What's in a phrase?*

[eyw04] *Gesture-based communication in human-computer interaction*, ch. Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library, Springer Verlag, 2004.

[FB04] E. Flety F. Bevilacqua, *Captation et analyse du mouvement pour l'interaction entre danse et musique*, actes des rencontres musicales pluridisciplinaires: le corps et la musique. grame, 2004.

[Gav99] D. M. Gavrila, *The visual analysis of human movement: A survey*, Computer Vision and Image Understanding: CVIU **73** (1999), no. 1, 82–98.

[HJLA01] Chil-Woo Lee Hyun-Ju Lee A1, Yong-Jae Lee A1, *Gesture classification and recognition using principal component analysis and hmm*, IEEE Pacific Rim Conference on Multimedia (Heung-Yeung Shum, Mark Liao, and Shih-Fu Chang, eds.), Lecture Notes in Computer Science, vol. 2195, Springer, 2001, pp. 756–763.

[JKA99] Q. Cai J. K. Aggarwal, *Human motion analysis: A review (1999)*, Computer Vision and Image Understanding: CVIU **73** (1999), no. 3, 428–440.

[KTP03] K. Kahol, P. Tripathi, and S. Panchanathan, *Gesture segmentation in complex motion sequences*, 2003.

[Mul] Remy Muller, *Human motion following system using hidden markov models*.

[Rab89] Lawrence R. Rabiner, *A tutorial on hidden markov models and selected applications in speech recognition*, Proceedings of the IEEE **77** (1989), no. 2, 257–286.

[Rin] The Laban Ring, *Notation du mouvement*.

[Sch04] Diemo Schwartz, *Data-driven concatenative sound synthesis*, Ph.D. thesis, Ircam, Paris, 2004.

[SDP00] F. Sparacino, G. Davenport, and A. Pentland, *Media in performance: interactive spaces for dance, theater, circus, and museum exhibits*, IBM Syst. J. **39** (2000), no. 3-4, 479–510.

[SEY$^+$] Asako Soga, Mamoru Endo, Takami Yasuda, Bin Umino, and Takaaki Kaiga, *Web3d dance research project*.

[Tro02] N. F. Troje, *Decomposing biological motion: A framework for analysis and synthesis of human gait patterns*, Journal of Vision, vol. 2, 2002, pp. 371–387.

[Tur] Matthew Turk, *Gesture recognition*.

[Vol03] Gualtiero Volpe, *Computational models of expressive gesture in multimedia systems*, Ph.D. thesis, InfoMus Lab, DIST – University of Genova, 2003.