

1 Sommario

Questa tesi affronta il problema della segmentazione del movimento umano.

In particolare si focalizza su come un osservatore identifica e separa in fasi una danza. Inoltre nella tesi sono studiati e sviluppati algoritmi che emulano questo comportamento in modo automatico. Come passo preliminare sono stati condotti alcuni studi sulla segmentazione basata sull'identificazione di fasi di pausa e di movimento. A validazione degli algoritmi sviluppati è stato condotto un esperimento su alcuni soggetti ai quali è stato chiesto di eseguire manualmente la segmentazione mentre esaminano video di danza. Risultati e raffronti sono riportati nel capitolo 8.

Sono state inoltre condotte ricerche sugli atteggiamenti posturali allo scopo di evidenziare se e come la postura possa influenzare l'esecuzione di movimenti e, di conseguenza, la loro segmentazione.

Lo studio può essere riassunto in tre fasi: (i) individuazione di parametri adatti ad effettuare la segmentazione, (ii) sviluppo di algoritmi per segmentare (basati sui parametri individuati), (iii) validazione degli algoritmi.

Il primo parametro preso in considerazione è stato l'equilibrio (sia statico che dinamico). Fonti di ispirazione variano dalla biomeccanica a teorie umanistiche, ad esempio nella coreografia. La tecnica basata sulla valutazione di quello che abbiamo chiamato equilibrio statico è risultata particolarmente utile nell'identificazione di movimenti periodici che coinvolgano i piedi (es. la deambulazione o la corsa) e nella loro suddivisione in fasi elementari. L'equilibrio dinamico (cioè lo studio dell'equilibrio utilizzando caratteristiche dinamiche del moto in diverse finestre temporali) è stato oggetto di analisi al fine di valutare stati di equilibrio raggiungibili in seguito a decelerazioni. Lo studio in questa direzione ha permesso di ottenere risultati preliminari, come discusso nel capitolo 6.4. Sono state prese in considerazione varie caratteristiche del movimento, tutte, comunque, estratte da segnali provenienti da una sola telecamera fissa che riprende un solo ballerino: il PAD (percentuale di accelerazioni e decelerazioni), gli zero-crossing di accelerazioni e decelerazioni (Zhao, 2001; Bindigavanale, 2001), gli zero-crossing delle componenti di velocità dell'arto che guida il movimento nella danza in esame. Quest'ultima caratteristica, sotto determinate ipotesi (riportate nel capitolo 3), può essere utilizzata per una segmentazione del moto che sia basata sui cambi di direzione seguiti dal danzatore come fattore discriminante per separare in fasi la danza. Gli algoritmi sviluppati sono stati implementati come moduli software per la piattaforma EyesWeb (sono perciò stati codificati in linguaggio C++, all'interno del sistema operativo Microsoft Windows®). La validazione degli algoritmi è stata condotta utilizzando video sia dell'archivio del Laboratorio InfoMus (ballerino e coreografo, Giovanni Di Cicco) sia di quello di Karlsruhe (ZKM): in questo caso sono stati utilizzati due frammenti di danza contemporanea su coreografie di William Forsythe.

1 Summary

This thesis faces the problem of segmentation of human movement. In particular, it focuses on investigating how observers identify "phrases" in dance performances, and on the development of computer algorithms to emulate such behaviour (Camurri et al., 2004) based on a single videocamera signal.

As a preliminary step, some work has been carried out to segment movements in motion and pause phases, using algorithms based on the Quantity of Motion (QoM) (Trocca, 2001, Mazzarino, 2002, Volpe, 2003). An experiment with subjects performing the same segmentation task was carried out to validate this algorithm, whose results are described in chapter 8.

Some research about posture was also carried out to evaluate if and how it can influence the execution of movements, and how it determines their segmentation. In this scenario, my work consisted in finding further algorithms for performing segmentation.

The study can be summarized in three steps: (i) individuating motion parameters that can be suitable for the segmentation task, (ii) developing algorithms for segmentation based on the identified parameters, (iii) validating the algorithms on a reference archive of movements with a particular focus on dance performances.

The first motion parameter I took into account has been Equilibrium, both static and dynamic, using different sources, from biomechanics to choreography. Static equilibrium techniques resulted particularly useful in identifying periodic movements involving feet (like walking, running, etc.) into strokes. Dynamic equilibrium (in our study, the measure of equilibrium using dynamic cues in different time windows) has been investigated to evaluate equilibrium states reached from decelerating preceding phases. The study in this direction only gave preliminary results, as discussed in chapter 6.4.

Several motion parameters have been considered, all extracted from a single (fixed) videocamera signal of one dancer on stage, such as PAD (percentage of accelerations and decelerations), zero-crossings of accelerations and decelerations (Zhao, 2001; Bindigavanale, 2001), zero crossings of the components of velocity of the "guiding limb" in the dance. This feature, under certain hypotheses (see chapter 3), can be used for segmentation, using the changes of direction followed by the dancer.

The developed algorithms have been implemented as software modules for the EyesWeb open architecture (coded in Visual C++ language under Microsoft Windows operating system). The validation of the algorithms has been conducted using videos recorded in our Laboratory (dancer and choreographer Giovanni Di Cicco), and from an archive from ZKM (Karlsruhe) on Forsythe's ballet fragments on contemporary dance.

2 Introduction

The work described in this thesis is in the framework of a more general project on modelling and analysis of non-verbal communication with a particular focus on expressiveness in full-body human movement. In particular, we focused on the development of automated techniques and real-time algorithms for human movement segmentation.

The problem faced in this thesis originated from the more general problem of segmenting human movement into coherent “units”, both from the point of view of the performer of the movement and from the point of view of any observer.

With motion segmentation, in fact, we mean the division of movements’ streams into phases according to some features or criteria perceived by observers, rather than the ‘separation’ of a human body with respect to the background of which it is a part.

Many scenarios appear to have links and connections with our work, which is included in the context of multimedia content analysis. In particular the work can take great advantage from a cross-disciplinary approach and it can highly benefit of cross-fertilization among scientific and technical knowledge on the one side, and art and humanities on the other side.

This need of cross-fertilization opens novel frontiers to research in both fields: if from one hand scientific and technological research can benefit of models and theories borrowed from psychology (e.g., Krumhansl’s studies (1997)), social science, art (music and dance, in particular) and humanities, on the other hand these disciplines can take advantage of the tools technology is able to provide for their own research, i.e., for investigating the hidden subtleties of human behaviour at a depth that has never been reached before. (Camurri, Mazzarino, Menocci, Rocca, Vallone and Volpe, 2004)

Nowadays the relevance of movement and gesture as a main channel of non-verbal communication is becoming evident, and a growing number of researches are developing in this direction (e.g., see the Gesture Workshop series of conferences started in 1996).

From a cross-disciplinary perspective, research on expressive gesture descriptors can be built on several bases, ranging from biomechanics, to psychology, to theories coming from performing arts.

For example, in our work we have considered theories from dance and choreography such as Rudolf Laban’s Theory of Effort (Laban, 1947, 1963), theories from music and composition (after all, movements in dance performances are orchestrated just like notes in music), works by psychologists on non-verbal communication in general (e.g., Argyle, 1980), on expressive cues in human full-body movement (e.g., Boone and Cunningham, 1998; Wallbott, 1980; Krumhansl, 2003), biomechanical works on human body motion, etc.

A special focus, anyway, has been devoted on expressive gesture in dance.

In particular, we have studied and integrated two relevant domains: that of dance and choreography together with that of computer vision and motion analysis techniques.

In fact, humanistic theories from dance and choreography, such as the theory of Effort by Rudolf Laban¹, explain – from the artist’s perspective – the non-verbal gesture language within human movement.

In order to be effective, the approaches have to start from a quite constrained framework where expressiveness can be exploited to its maximum extent. One such scenario has been found in dance and it is also a good example for carrying out a task like that of segmentation.

Dancers and actors are, in most cases, aware of techniques that can be used to emphasize some movements or gestures and they are also able to evoke, with their actions, particular emotions or moods, using them at will to convey expressive contents to the audience. Expressive content concerns aspects related to feeling, affect and intensity of emotional experience. For example, the same action can be performed in several ways, by stressing different qualities of movement.

In this manner it is possible to recognize a person from the way he/she walks, but it is also possible to get information about the emotional state of a person just by looking at his/her gait, e.g., if he/she is angry, sad, happy (as Pollick stressed in his papers). (Pollick, June 2001, August 2001)

In cases of gait analysis, we can therefore distinguish among several objectives and layers of analysis: a first one aiming at describing the physical features of movement, for example in order to classify it, a second one aiming at extracting the expressive content that gait conveys², e.g., in terms of information about the emotional state the walker communicates through his/her way of walking. (Boone and Cunningham, 1998; Pollick, 2001)

For what concerns human motion analysis in its most general aspects, this field is nowadays receiving increasing attention (from an engineering point of view) by computer vision researchers. The interest is motivated by a wide spectrum of applications, such as athletic performance analysis (a), surveillance (b), content-based image storage and retrieval (c), video conferencing (d), etc.

Aggarwal and Cai (1997) give an overview of the various tasks involved in motion analysis of the human body, as specified below.

(a) Segmenting³ the parts of the human body in an image, tracking the movement of joints over an image sequence, and recovering the underlying 3D body structure is

¹ Details are contained in the paragraph titled “The dance field”.

² Once some meaningful cues are identified, they need to be measured (possibly in real-time) on the expressive gestures the user performs.

³ In this case, “segmentation” is used with a different meaning from which that will emerge throughout this dissertation.

particularly useful for analysis of athletic performances as well as medical diagnostics.

(b) The capability to automatically monitor human activities using computers in security-sensitive areas such as airports, borders, and building lobbies is of great interest to the police and military.

(c) With the development of digital libraries, the ability to interpret video sequences in an automatic way will save tremendous human effort in sorting and retrieving images or sequences of them using content-based queries.

(d) Another kind of multimedia application includes video conferencing, whose pros and cons, in case of segmentation task, will be underlined later on.

In our study the methodologies for automatic segmentation of human movements have been analysed with a main focus on multimedia content analysis as a central application field: in the following pages we report our recent approaches and the various techniques employed to accomplish the task starting from sequences of images about dancers.

In order to find the most applicable and pertinent techniques we have first carried out a research about the existing methods, in order to have a clear idea of the state-of-the-art in the area probed by our work.

The review has been conducted with the aim to compare and judge the various studies existing in literature, highlighting (i) why we have taken into account some methodologies while having rejected others and, for those which resulted close to ours, (ii) the cases they can be applied to, together with (iii) their main advantages and drawbacks.

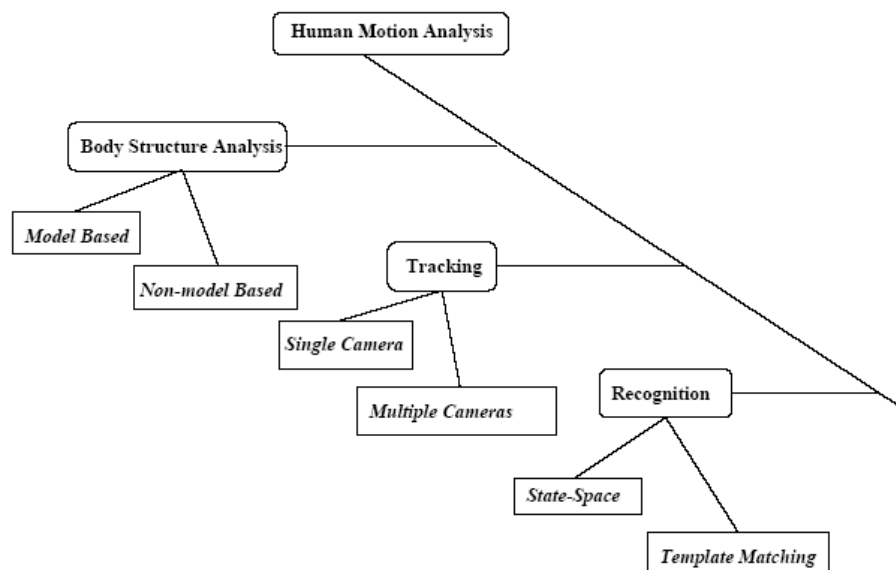


Figure 2.1: the three big areas of human motion analysis addressed by the Aggarwal and Cai (1997).

The figure depicted in the previous page shows the three areas on which human motion analysis mainly concentrates, according to Aggarwal and Cai's point of view (1997): body structure analysis (developed mainly by biomechanics), tracking (for example for video conferencing scenarios, but also addressed by multimedia content analysis) and recognition (useful, for example, in the video-surveillance field and also in case of segmentation based on recognition of single gestures).

In (Aggarwal and Cai, 1997) the two authors also point out that, talking about motion analysis, there is always a trade-off between feature complexity and tracking efficiency: lower level features, such as points, are easier to extract, but relatively more difficult to track than higher-level features such as blobs and 3D volumes. This has been confirmed by our work: we have concentrated on tracking of points corresponding to joints or to the Centre of Mass of the body and the algorithm employed for motion segmentation needs precise tracking of the chosen points.

This is the reason why an accurate manual tracking has sometimes been necessary: to obtain reliable values of position. We can certainly affirm that getting consistent values has been one of the bottlenecks of our approach to segmentation.

Another possible way to approach the issue of motion segmentation might be through considering motion recognition: recognizing a certain movement allows its distinction among others in a flow of different actions: therefore we may have segmentation based on recognition of certain moves even if they are meshed with other unrecognisable movements.

Two typical approaches to motion recognition are addressed in the publication by Aggarwal and Cai (1997): that based on template matching some given images to pre-stored patterns (the preferred approach by Aggarwal, who used it in (Aggarwal and Ali, 2001)) and that based on a state-space models.

To say the truth, we have not used the first approach neither the second, since ours is based on motion segmentation conceived as a step before or even disconnected with motion recognition. We have tried to make motion segmentation without any recognition of the movements.

Our approach is rather an attempt to divide streams of dance movements in phases according to some kinematical features, with a particular focus on how observers would execute such a task. In fact, the basic units of movement we can detect (e.g., in a dance) and cluster together can be considered and analysed under three perspectives, sometimes connected one to another: that of the **performer** (it means that attention is focused on the physical execution of the movement and expressive gestures⁴ represent items of acted moves), that of the **observer** (it is based on the perception of movements) and finally that of the **choreographer**. This topic is detailed in the paragraph "The dance field".

⁴ The concept of expressive gesture is deepened in the paragraph "Experimental psychology".

It is to say that this distinction should not to be considered in rigid terms: that is, people who study these mechanisms must necessary be acquainted with all the three viewpoints, and models of gestures have actually been studied by paying deep attention to the three perspectives.

In our work the attention has been principally focused on gestures as units of perceived movement: we have been interested in studying algorithms able to extrapolate the same (o similar) phases that an observer would individuate.

From the observer's point of view, segmentation of movements often refers to major segments of time that people usually describe by single action verbs⁵.

Human activity is a continuous flow of single human action primitives in succession, just as dance is a continuous occurring in sequences of different "atomic" steps and movements, sometimes modified by transitions to the neighbouring steps⁶.

When humans move from one type of movement to another, they do so smoothly: in general transitions are not well defined, there is no clear beginning or end of a movement; therefore the detection of shifts between them is crucial and this aspect makes segmentation difficult.

This dissertation is introduced in a field already explored by many researchers (not only engineers) and touches an actual matter, still open, without a universal solution: it focuses on the development of paradigms and algorithms to create techniques able to segment movements automatically, to divide them in elementary events, in movement strokes.

Having to deal with motion segmentation a spontaneous question may be: *how* can motion streams be divided? According to actions, to gestures, to step dances (if dealing with dance performances) or what more?

So, before introducing the discussion about the employed techniques and the choice of the main application scenario of our work (that of multimedia content analysis), an interlude about the differences among action, movement and gesture has been deemed worthwhile (that is the reason why an entire paragraph, titled "Experimental psychology", will be devoted to this aspect).

Here we just want to highlight that, although many experiments about expressive gesture aim at individuating which motion cues are mostly involved in conveying the dancer's expressive intentions to the audience (during a dance performance) and measuring them in order to classify dance gestures and steps in terms of basic emotions, we have instead considered gestures by virtue of their physical characteristics of execution (without any respect to emotional mechanisms) and we have found that motion phases are similar to those commonly used in case of investigation of complex and tangled sound patterns⁷.

⁵ Differences among actions, movements and gestures are highlighted in the paragraph titled "Experimental psychology".

⁶ Analogous considerations are valid for co-articulation in speech.

⁷ See the paragraph titled "Analogies between motion and music" for details.

These phases are preparation, attack, decay, sustain, release and overshoot (Mazzarino, 2003). Laban, talking about gesture, considers three phases: preparation, stroke (execution) and retraction (of the arms back close to the body) (Zhao, 2001). The temporal duration of every phase varies according to its execution: the more difficult or more precise the execution of a certain movement, the longer its arrangement (so the duration of the preparation phase).

Posture influences all the three phases (Bindigavanale and Badler, 1998) and an entire paragraph is devoted to this issue, since postural attitudes influence our way to move, act and also to be steady.

As in Camurri, Lagerlöf and Volpe's study our work is focused on dance: in their paper (2002) the three authors approach the genre of modern dance to find a nonpropositional style of movements (since the basis of natural body expression can be developed in terms of dance movements).

We dealt with videos about choreographies of contemporary dance.

A spontaneous question may be: why a dancer? Why not a simple person moving and doing normal daily activities?

Our attention has been focused on dance, since it is an artistic expression of human gesture and the field of dance is enriched by many facets⁸.

As already stated, dancers and actors are, in most cases, taught the techniques that can be used to evoke an emotion and they are able to amplify some gestures, postures or defects.

They are also able to emphasize some movements or pauses. In fact, not only motion is important in expressive content communication: pauses also play an overriding role. During them the body can assume a particular posture and body postures are often considered as expressive gestures with a relevant function in conveying expressive content to the audience (Argyle, 1980).

Moreover, we have tried to focus our work on the analysis of motion regarding the whole body, not just a few joints: literature, instead, is full of researchers who took into account just parts of the body (Cutler and Turk, 1998; Zobl, Wallhoff and Rigoll, 2003; Masoud and Papanikolopoulos, 2003).

A main aim for which we have focused on segmentation is that, dealing with movies, it may sometimes be an impossible or heavy task to extrapolate the necessary features without dividing the long stream of moves in minor parts.

Within each motion segment, semantic and style information may, then, be extracted. The problem is: which feature is best suited for motion segmentation? Will it allow segmenting according to the same criteria that a human observer (who, normally, notices the changes of actions looking, most of all, at which are the limbs moving, at how much they are moving, etc.) would use? Should it be just one feature or more, according to the type of video and to the performance?

⁸ More details should be taken into account talking about dance: refer to the paragraph titled "The dance field" for more information.

First of all, we have searched for features (one or more) carrying general information, whose utilization is possible for every kind of moving person (from the agilest dancer to any patient with motor disease). Kinematical cues accomplish such aim.

The methods we have tried to employ aim at being independent from the properties of the body silhouette: that is, we have only concentrated on kinematical cues (positions, velocities, accelerations and some quantities derived from them), features that are general and valuable for every kind of person, related just indirectly to the size or shape of the silhouette. Obviously also kinematical cues depend on the body mass (i.e., a heavy person might have lower values of velocity than a thinner one), but the dependence is indirect: that is, none of the calculated cues depends on the weight or height of the person for which we have evaluated it (this is especially true if the changes of scale are little, because otherwise we would need a normalization).

This thesis will not be presented as a dissertation divided into parts with the state-of-the-art as opposed to the new techniques we have introduced: we will rather report the various and different techniques we took into account, highlighting our approaches as opposed to “other methods”⁹, for which we will underline the cases in which they can be applied and when it is not possible.

As often happens, our research has started from naïve hypotheses and the very first results were quite ambiguous and unclear... probably because of the videos initially used (Di Cicco¹⁰) and because of the bibliographic references we have taken into account¹¹.

For this reason, afterwards, we have constrained the scenario and considered other videos (Forsythe’s cd-rom¹²).

As already stated, comparisons with other possible techniques, even with different perspectives, have been really useful: some of them have constituted the starting point of our study, some should represent a direction for future potential researches since there is not a unique way to proceed in the field of segmentation of human movement.

⁹ We have dedicated some paragraphs to the other techniques found in literature and applicable in some circumstances and for purposes similar to ours.

¹⁰ In these videos the dancer uses, for his moves, all the stage, the foreground as the background: this aspect makes the extraction of features difficult and unreliable for reasons highlighted in the third chapter. Other researches, instead, worked with videos in which there is only a lateral viewpoint of the body moving (this is a big limitation).

¹¹ Some techniques we read about seemed to be applicable to our videos for our purposes, but we have discovered later that our results were really different from those obtained by other researchers, maybe because of different video resolution.

¹² See Chapter 3 for details about them.