Recognizing Classical Ballet Steps Using Phase Space Constraints

by

Lee Winston Campbell

B.A., Physics
Middlebury College, Middlebury VT
June 1978

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN MEDIA ARTS AND SCIENCES
at the
Massachusetts Institute of Technology
February 1995

Signature of Author _____
Program in Media Arts and Sciences
6 September 1994

Certified by _____
Aaron F. Bobick
Assistant Professor of Computational Vision
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by _____
Stephen A. Benton
Chairperson
Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Recognizing Classical Ballet Steps Using Phase Space Constraints

by
Lee Winston Campbell

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on September 6, 1994
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

Abstract

This work deals with the problem of representing and recognizing human body movements, given XYZ tracking data. Prior approaches by other researchers used a smaller number of classification categories, which demanded less attention to representation.

We develop techniques for representation of movements based on space curves in subspaces of a "phase space." The phase space has axes of joint angles and torso location and attitude, and the subspaces are subsets of the axes of the phase space. Using this representation we develop a system for learning new movements from ground truth data by searching for subspaces in which the movement to be learned describes a curve which is easily separated from other movements. We then use the learned representation for recognizing movements in data.

We train and test the system on nine fundamental movements from classical ballet by two dancers, and show the system can learn, recognize and segment out five of the movements accurately, but confuses one pair of movements from one dancer and another pair from the other dancer. Finally, we suggest how the system can be improved and extended.

Thesis Supervisor: Aaron F. Bobick
Title: Assistant Professor of Computational Vision, MIT Media Lab

This work was supported in part by a grant from Interval Research.

Recognizing Classical Ballet Steps Using Phase Space Constraints

by

Lee Winston Campbell

The following people served as readers for this thesis:

Reader: _____

Richard A. Bolt
Senior Research Scientist, Advanced Human Interface Group
Program in Media Arts and Sciences

Reader: _____

Jessica K. Hodgins
Assistant Professor, College of Computing
Georgia Institute of Technology

I'd like to thank my advisor, Aaron Bobick, for originally suggesting this fascinating topic, for his ideas and suggestions while I was trying out approaches, and for his useful comments and insights while writing up the thesis. The two years that have gone by so quickly have been rewarding, interesting, and fun, due in no small part to Aaron.

The readers for this thesis, Dick Bolt and Jessica Hodgins, were helpful both in clarifying my thesis proposal and in preparing the final document. Ms Hodgins was particularly insightful in her comments on rough drafts.

Dean Wormell, of Adaptive Optics Associates, has been helpful above and beyond the call of duty in loaning us AOA's tracking cameras and processors, helping to gather the movement data, and teaching me to use the tracking software. Dean and AOA have been a pleasure to collaborate with.

I've been aided by many interesting conversations I've had with the other students in the Vision and Modeling group; particularly Stephen Intille, Nassir Navab, and my officemate Andy Wilson. VisMod has provided an exciting, stimulating environment and its people have been great resources.

Our ballet experts and dancers, Katy Evanco, Meg MaGuire and Chen Yu have been enthusiastic, patient, and helpful; even while dancing with reflectors attached to them and while answering numerous questions about the art and practice of ballet.

Finally, my brother and sister, Bruce and Jan Campbell, and my parents, Malcolm and Jeanne Campbell, have provided their steady love and support at all times. I have no adequate way to thank my parents for all they have given me over the years except to acknowledge how they have encouraged and helped me in every endeavor.

Contents

Chapter 1

Introduction

1.1 Computers Understanding Human Body Motion

As computers become more inexpensive, powerful, and ubiquitous, they will take on more of the drudgery in people's lives and watch out more for our safety. They may watch traffic for pedestrians in crosswalks, monitor other traffic from cars, and observe operators of dangerous machinery. In order to do these things, computers must see and understand people and the movements of their bodies. This thesis is a small step towards the goal of making computers into guardians of our safety and more useful helpers.

To understand the movements of people means much more than merely quantifying the details of the movements — at the very least it means identifying or recognizing the movements, and equating them with symbols. This problem is analogous to the problem of understanding speech: in each case the input is a signal represented by a large amount of numerical data and the output is a sequence of symbols expressed at a level of detail which is the natural level of detail for human reasoning and speaking.

In order to study a problem it is useful to work in a well–defined domain where the scope of the problem can be controlled. In order to study the identification and recognition human movement, the domain must have the following three features:

A vocabulary of symbols is important because symbols are equivalent to categories for classification. Using symbols or categories from an existing language means we are provided with a ready-made and tested set of classification categories. The symbols exist because they have been found useful for classification.

Level of detail of the symbols is important because if computers are to reason and communicate in human ways, they must do so at the same level of detail as humans. Speech understanding is much more important than sound characterization because speech symbols are a major channel of human reasoning and communication. Similarly, if movements can be understood by computers at the same level of detail as notions such as walking, sitting down, reaching, and falling, then computers will have some of the right symbols with which to reason about human safety, needs, and intentions.

Availability of ground truth for the movement identification is important because any system which purports to do recognition must be tested, and the ready availability of ground truth data makes it easy design tests. Ground truth can also be used by a system that automatically "learns" movements from training data (supervised learning).

Ballet is a good testbed for work in understanding human body movement because it has the requisite features: there is a vocabulary of some 800 names of steps as well as several notation languages; the vocabulary has been useful in ballet for over a hundred years, thus it is at the appropriate level of detail for human reasoning and communication; and human observers can easily provide ground truth identifications against which to check the computer recognition.

1.2 Goal of this work

This work is a step towards a system capable of seeing and understanding human body movement. This is a huge project with many tasks, and many computer vision researchers are working on parts of it. One part of the task is finding a person in imagery and identifying their body pose. In its most general form this problem is extremely difficult: people are non-rigid, articulated, self-occluding objects, all of which greatly complicates the problem of finding them in imagery. Many researchers are working on finding people in imagery assuming a stationary camera, specially marked clothing, or some other constraint. We have chosen to assume solutions will be found to the problems of finding and tracking people and recovering their 3D body pose, so the system developed for this thesis takes as input XYZ tracking data instead of imagery. This leaves us free to concentrate on the

problem of recognition of movement. An important part of the recognition system is a representation for movement which facilitates recognition.

Thus the goal of this work is to develop a system capable of representing and recognizing classical ballet steps, given an input of XYZ tracking data. This is a subgoal in the larger task of understanding human body motion from images. However, even given XYZ tracking data, there are significant challenges: How does one capture what is invariant in a ballet step while remaining tolerant of the normal variation between dancers? What features can be used, and how can they be extracted from the data? How shall dance steps be represented? All these questions are interrelated – no single one can be answered without addressing them all. We shall approach the problem by beginning with representation issues and then developing recognition techniques.

Classical ballet comprises some 800 steps, depending on how they are counted. Major categories of steps include jumps, turns, traveling steps, and stationary poses or attitudes. There also exists a set of movements performed frequently in practice which make up subparts of many of the other movements. We will do recognition experiments in this set of movements, as will be explained in Section 2.3.

1.3 Other Applications

The goal of realtime general understanding of human body movement may be twenty or more years in the future. However, understanding limited to a specialized domain, or offline, slower-than-real-time understanding of human body motion may be realized sooner, and there are applications of it which have social and financial value.

One such application is video annotation [10]. Entertainment companies, newscasters, and sports teams are acquiring ever larger video databases. If these can be automatically scanned and annotated, the video can be organized, catalogued, cross-referenced, and searched for keywords. Powerful text search tools can be applied once the mass of signals has been annotated with natural language symbols, and speed of annotation is relatively unimportant.

Another interesting application is fair judging of athletic events such as gymnastics, diving, and ice skating. In such an application the task of the computer would not be recognition so much as identifying the deviations from ideal form. An automatic judge

would also be useful as a training tool, helping to teach athletes specifically what to change to improve form and performance (already there are rudimentary automatic systems to diagnose golf swings).

Finally, automatic notation of dance and movement may be a useful application. Choreography has been called "the throwaway art" because the choreography was lost after a company stopped performing a particular version of a work. Systems of dance notation and general body movement notation such as Labanotation were developed to record choreography. This has been supplanted to some extent by videotape, but notation can convey the choreographer's intent in a way that videotape cannot. Moreover, Badler [5] proposed an animation and simulation system which could take several forms of input and output, and used Labanotation as its internal representation. Motion capture for animation is growing in commercial importance, and a system capable of changing the expressive content of the motions by re-generating the motion from notation would be a useful tool.

1.4 Outline of the Thesis

Having provided the motivation behind this work, we must now explain our approach to it.

Chapter 2 briefly describes the domain of ballet as it pertains to this thesis, and then states the detailed problem addressed by this work. Chapter 3 lays out the assumptions behind the problem and surveys the literature related to the assumptions and problem. Chapter 4 develops a phase space approach to representing and recognizing human body motion. Chapter 5 describes the results of testing, and Chapter 6 sums up the work thus far and areas needing improvement in future work.

Chapter 2

Problem Description

The problem of understanding human body motion from images is one that leads into such diverse areas as dynamics, athletics, and cognitive science. We will briefly survey several areas of human motion study and then discuss the nature of balletic motion. We will then state the specific problem this thesis addresses.

2.1 Human Body Motion

Human body motion has many constraints. Some of the constraints come from the laws of physics. Others, which come from the rules of an athletic or artistic form, we will call cultural constraints. To understand the motions, it is first necessary to understand the constraints. These can then be exploited to guide search and limit computation.

For example, gymnastics routines often contain many airborne rotating maneuvers. Without ground contact, the center of mass travels in a parabola with constant downward acceleration and the angular momentum is constant. Within these constraints, gymnastic skill is expressed in the height achieved, the number of rotations, control of twisting and tumbling, and especially control of body form. [1] It might seem that an airborne person has no freedom to alter his or her trajectory, but gymnasts adjust their moment of inertia by piking, arching and arm motions. These actions allow limited control of body attitude.

[1] There is an interesting reason why form is a cultural constraint. It is constrained because it is difficult to achieve, and it is difficult because many gymnastic maneuvers involve rotation about a line through the body from left to right. This is the axis of median moment of inertia and rotation about it is unstable for small errors in initial conditions – tumbling occurs. If the gymnast tumbles uncontrollably, she or he must break form to land safely. Thus the constraint of good form is prized because the subtle physics of rotational inertia make it difficult.

Gaits are constrained (or perhaps defined and categorized) to be rhythmic and repetitive, and these constraints allow searches for periodicity [33]. Both Hogg [25] and Rohr [38] treated walking as a 1DOF motion for purposes finding limb orientations on a walking person.

Hand gesture recognition has many similarities with whole body motion. The hand has many degrees of freedom (10 1D joints and 5 2D joints) and a great range of expression. However, it is subject to different constraints: it does not need to obey the constraint of maintaining balance and support which the whole body obeys. Its motions are highly context sensitive. The most extreme example is American Sign Language, which is a full context sensitive grammar plus all the puns, slang, abbreviations, and neologisms one expects from a living language. Moreover, the hands are usually involved in co-articulation, [11] either with the voice and/or the face and posture. So to understand hand gestures it is necessary to understand context and the other modes of articulation. This is a much more complex set of constraints than ballet.

2.2 The Domain of Ballet

Ballet has its own set of cultural and physical constraints. Classical ballet is made up of a finite number of discrete movements or steps. People with expertise can easily name and recognize most of them, even when seen on video. In addition to named steps there are classifications or categories of steps and written notation languages for recording ballet and other forms of dance.

There are three major schools of classical ballet: Russian, French, and Cecchetti or Italian. Since some steps are performed differently in different schools there are about 300 or 400 steps per school and about 800 to 1000 steps total. Major categories of steps that occur in performance include stationary poses, turns and pirouettes, linking steps and movements, jumps, and beats (a fluttering of the legs during a jump). Steps that occur in practice or class and during warmup include relevés, battements and ronds de jambe. Most of the steps can be started from any of five foot positions (Although first, fourth, and fifth positions are far more common). Many of the steps can be done either on pointes or on demi-pointes. There are 8 arm positions in the Cecchetti school, six in the French, and four in the Russian. There is more freedom for the arms to be expressive and to take a range of

positions, and not all arm positions occur with all steps [19, 46].

The practice steps are interesting because they tend to be simple movements with one degree of freedom which are pieces of the more complex movements. Thus they are a sort of syllable or phoneme of dance and we call them "atomic" movements. They include plié, relevé, tendu, développé, and battement. Many steps involve one support leg and the other leg is termed a "working" leg.

Although most steps can be categorized, there are also steps which do not bear any family resemblance to any other steps. Historically, they may have entered ballet from folk dances. Examples of such steps are pas de basque and pas de chat [50].

When steps occur in sequences, the beginnings and ends often change a little to accommodate the neighboring step. This is analogous to continuation in speech, where a vowel sound can be changed a great deal by context. Another source of variation is called "extension:" the term refers to the height to which a leg is lifted or the height of a jump. Because of variations in style and bodies, different dancers may do some movements with less extension; sometimes the choreographer may utilize the special abilities of a dancer and call for a "bravura" form of the movement with more extension.

Ballet is subject to several cultural constraints. Placement and technique are the terms for the correct positions of the limbs and ways of moving. They are learned one ballet step at a time. Yet some generalizations can be made about ballet's cultural constraints. An obvious one is called "turnout" and it requires that the toes and knees be pointed completely to the sides, 180° away from each other. Another constraint is grace: it is better to do a movement with less extension gracefully than with large extension that shows straining. Another constraint is maintenance of support and balance. Although that may seem too obvious to mention, we believe this constraint has a profound effect on whole body movement and constrains most of the motions of the major masses of the body to one intrinsic degree of freedom. As an example of the support and balance constraints outside of ballet, consider the movement of sitting down: people bend at the ankle, knee, hip and spine when they sit, and though there are many possible combinations of angles that will land them in a chair, most people sit with similar movements. It is likely this similarity occurs because people try to maintain their balance and comfortably support their torsos during the movement.

The most popular form of dance notation today is Labanotation [21], first published by

Rudolf von Laban [49] in 1928. Labanotators are certified by, and can be hired from the Dance Notation Bureau in New York. The reader may wonder why, with the low cost of videotape, anyone would wish to notate dance by hand. The answer lies in an analogy to music: sheet music of a solo conveys the intent of the composer and indicates what parts of the performance are left to the performer's interpretation, while recorded music only conveys a particular performance, complete with errors and the performer's interpretation. Choreographers would like something analogous to sheet music from which dancers could learn new repertoire.

Labanotation provides varying amounts of specificity. Positions can be prescribed all the way out to feet and fingertips, or some of the details can be left up to the dancer. There are ways to express coordinated movements or coordinated dancers, points of contact, and paths on stage as well as positions of the body. Thus Labanotation is an attempt to capture the constraints of choreography as well as the constraints of the movements. Unfortunately, notation is slow: on the order of an hour's work by a highly trained person for each minute of notated dance. At present only a few large professional dance companies notate the dances they commission.

Good reviews of movement notation are in [32, 20]. Other dance notation systems[2] in current use include Benesh [7], Eshkol-Wachman [15], and Sutton [44].

2.3 The Specific Problem

The specific problem this work attempts to solve is to learn and identify a set of nine atomic ballet movements from XYZ tracking data. The nine movements (see Figure 2-1), all performed starting from a flat-footed position with legs turned out and with the right leg working and the left leg supporting, are:

1. *plié* lowering the torso by bending the knees, then rising back up (used when launching and landing every two-legged jump, and with many other steps);

2. *relevé* rising up on the balls of the feet and then lowering (preparation for many steps; a plié or relevé is part of almost every ballet step);

[2] also indexed under kinesiology. See also gymnastics notation.

3. *tendu á la seconde* sliding the foot to the side and bending the ankle as needed to maintain toe contact with the floor (preparation for many steps);

4. *dégagé* raising the leg rapidly about 45° to the side (part of many steps involving transfer of weight or traveling),

5. *fondu* a plié on the supporting leg while the working leg bends at the knee and points the toe down (a fluid step frequently used by itself);

6. *frappé* raising the working foot vertically by bending the knee and hip until the hip makes a 45° angle, then rapidly straightening the knee and ankle to kick to the side (part of many jetés (one-legged jumps) and assemblés);

7. *développé* like frappé but the foot is raised until the hip makes a 90° angle before straightening the knee (frequently used before a promenade, or by itself, especially in partnering, to extend the leg out horizontally);

8. *grand battement á la seconde* with the knee straight, raising the leg to the side until the foot is at shoulder height;

9. *grand battement devant* same as battement á la seconde except the leg is raised to the front (both grand battements can be used by themselves in allegro passages instead of développé).

These steps were chosen for the following reasons:

- this gives leverage for recognizing more complex steps if several parts of the step are known;

- several of the movements have similarities which will test the system's ability to make fine distinctions;

- several of the movements are quite simple and thus suitable for beginning development of a representation.

The input data was recorded in two different sequences, one from each of two dancers. The task will be to learn the movements from the two sequences with a single dancer-independent representation for each movement, and then test the representation by recognizing the start time and duration of the movements in the same sequences. This is known

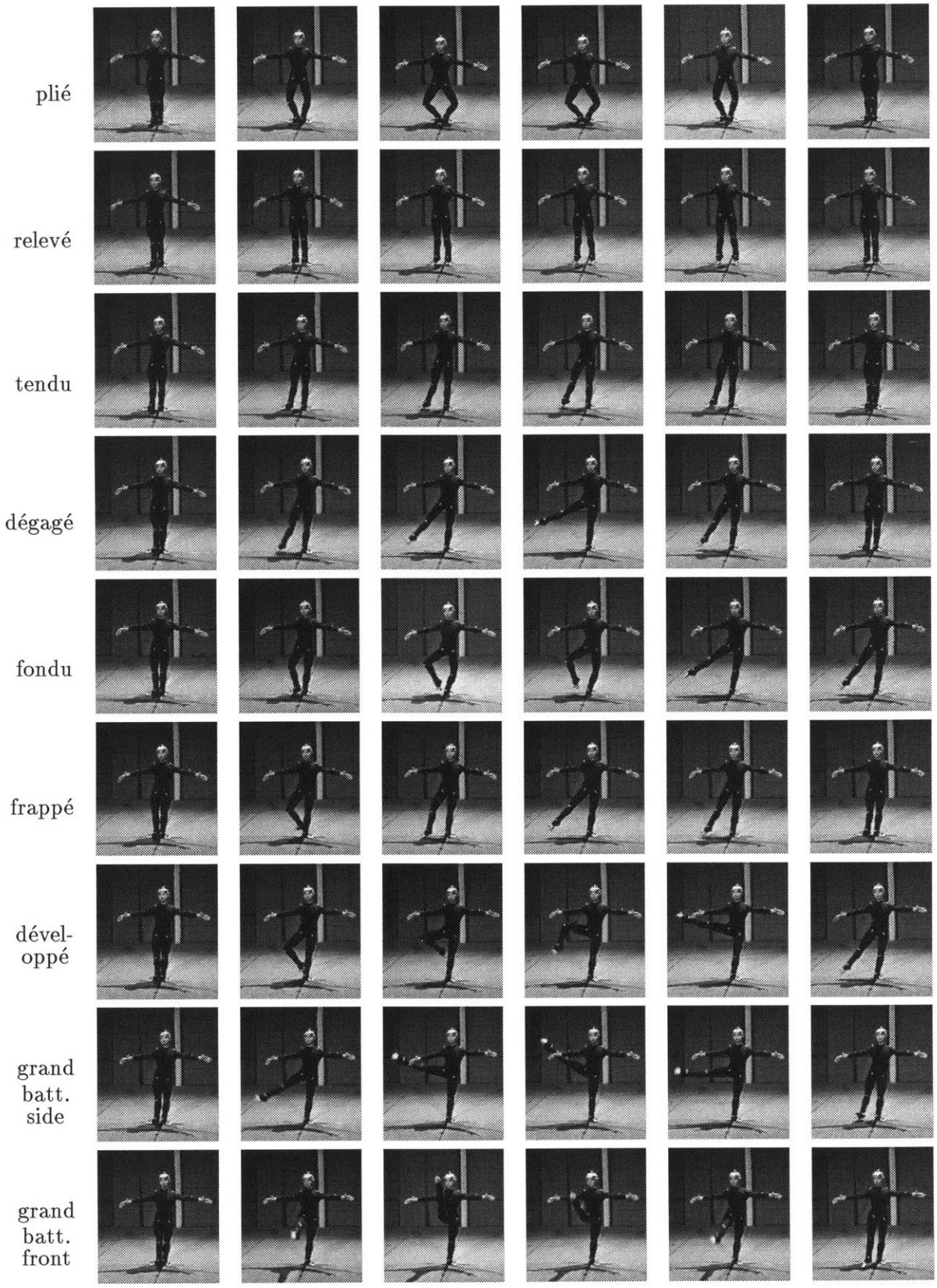

Figure 2-1: Nine "atomic" ballet moves.

as a resubstitution estimate [12]. The reasons for using this weaker form of test are given in Chapter 5. We will use two measures of success: the less rigorous one is to correctly and uniquely state that the movement has taken place, and the more rigorous one is to count the number of time steps that are correctly labeled and the number that are incorrectly labeled.

Chapter 3

Related Work

This thesis is a link in the chain to a larger goal: understanding human body motion in real image sequences. Because the whole task is as yet too challenging, we have chosen the subgoal of recognizing ballet steps given XYZ tracking data from points attached to the dancer's joints, and we have chosen a method which involves using a phase space representation of joint angles on a human body model. We make several assumptions in claiming that our subgoal is a part of the larger goal, and several more in our choice of representation. In this Chapter we make our assumptions explicit and then look to the literature for justification.

Our choice of subgoal makes the following assumptions:

1. that the task of determining dance steps from tracking data is achievable;
2. that recognizing human body motion from real imagery is too difficult a task;
3. that it is an acceptable simplification to use input from tracking data instead of images;
4. that there exist low-level vision algorithms which could derive the necessary input from images.

Moreover, in our choice of representation we make these assumptions:

5. that a human body model is a necessary part of the representation;
6. that the phase space method and representation (described in Chapter 4) is appropriate;

7. that other representations cited in the literature are not well suited to this task.

The discussion of moving light displays justifies assumptions 1, 2, and 4. The discussion of spatio-temporal behavior without models justifies assumption 5. The discussion of phase space justifies assumption 7; and the discussion of human body pose recovery supports assumptions 3 and 7.

3.1 Moving Light Displays

Psychophysics

The perceptual experiments of Johansson in 1973 [28] opened up the field variously known as "Moving Light Displays" or "Point Light Walker Displays" or "Biological Motion" (referring to patterns of movement created by living creatures). Johansson attached reflective patches to ankle, knee, hip, shoulder, elbow, and wrist joints, and to the head, and took videotapes with the signal adjusted such that only white spots on a dark background could be seen. In some tests, subjects viewing the videos were able to tell whether or not the motion was biological in times as short as 100 or 200msec [29]. In other tests, common motions were added to or subtracted from all the points, and subjects still identified the motion. In most of the tests 100% of the subjects responded correctly, even though they had not been prepared or trained in any way. Another experimenter [13] found that other activities were recognized such as hammering, box lifting, ball bouncing, and stirring; and two-person activities such as dancing, greeting and boxing. Yet another series of experiments [6] showed that people could determine the gender of a walker and even identify a friend based on MLD's of their walk.

It is perhaps surprising that people can make such quick identifications from such a paucity of data. The effect is quite robust. For computers, the paucity of data may be a benefit because they have such difficulty identifying and tracking objects in real images. Since people can identify movements in MLD so easily, perhaps it is possible for a computer to make such identifications as well. Johansson's work raises the question of whether people viewing MLD's identify structure first and then identify parts (e.g. hands and feet) from the structure; or whether the motion leads us to identify the parts first, after which we infer the structure.

Computer Vision

Johansson's experiments show that the model of the human body as moving locations of joints rigidly linked by limbs is in some sense a sufficient model for a number of recognition tasks. This has inspired work in the computer vision research community to explain how details of body pose can be extracted from the 2D projection which is a moving light display.

Though not necessarily inspired by Johansson, Ullman proved a basic structure from motion theorem: Given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views is uniquely determined [45]. This theorem plus Ullman's rigidity assumption and rigidity test suggests that the rigid links of a jointed body can be found.

Rashid [37] made an early attempt to link points into a structure based on clustering of 2D positions and velocities. His system tracked points through frames, grouping and linking them into objects according to their relative velocities. The system was tested on synthetic 2-D perspective projection MLD's of several objects such as a man walking a dog. The points that were rigidly linked in space as well as independently moving sub-parts could frequently be found by the program even though their projections were not rigidly linked. Although velocity clustering may provide a good first estimate of how points are linked into objects, it is too weak a model of objects to be relied on.

Webb and Aggarwal [51] assume the links between points are known, and that points rotate around fixed axes. They use orthographic projection and show that points will sweep out ellipses in the projection, and that the eccentricity and orientation of the ellipses determine the 3-D axis of rotation to within a reflection. Their work was later extended by Asada [4] (with an assumption of constant angular velocity) to a case where, for example, points rotate about the shoulder which in turn rotates about some other axis.

Hoffman and Flinchbaugh [24] present a method which uses the fixed axis assumption of Webb, but recovers both linkages and axes. Thus they recover 3-D structure in the case where the axis directions are fixed. This is a good approximation for the cases of walking and running, but not dancing.

Zhao [53] built a system to do accurate automatic 3D model fitting to 2D moving light displays. It takes as input target centroids representing a monocular 2D sequence of a moving light display, plus initial identifications and camera calibration parameters. It uses

Jack[1] as a human body model. It tracks points, recovers joint angles, and corrects limb lengths on the model. It was tested with data that had been filtered through *Jack* in the following way: Real world data was collected for sitting people reaching; the *Jack* model was constrained to follow the data as closely as possible and extract joint angles, and then the joint angles were played through another *Jack* model projected to 2D to make synthetic tracking data. Zhao's tracking system took that data as input, as well as a body model with errors in starting limb lengths. Operating on the test data, the system demonstrated near perfect recovery of the information lost during monocular imaging. However, it is not clear how robust the system is to errors in tracking point placement; in the tests, it appeared the tracking points were placed in the centers of the actual joints, which is not possible with real people.

Each of these systems makes similar assumptions about the input: that points or features corresponding to limbs can be found and tracked through successive frames to provide an MLD input. Webb and Zhao assume starting identification (i.e. linkage) is given; Webb and Flinchbaugh assume fixed axes. This body of work, taken as a whole, shows that it is within grasp of computer vision to recover structure and thus joint angles from a moving light display or its equivalent. This justifies our use of 3D structure data as input to our system for dance step recognition.

3.2 Spatiotemporal Behavior Without Models

The following is work aimed at doing recognition directly from motion, without the intermediate stage of recovering structure.

Rangarajan, Allan and Shah [36] state a goal of distinguishing between two similar objects with different motions, or two objects with the same motion but different shapes. They developed a system to match trajectories. They represent a trajectory as two curves: one is speed vs time; the other is direction vs time. They then produce scale space images of the two curves, and run a matching algorithm on the scale space zero crossings. They used as test input a trajectory, three versions of that trajectory that were rotated, magnified, and noise added, and a distinct trajectory. They show that the more similar trajectories

[1] a commercial package for simulating human body movements, developed by Badler at University of Pennsylvania

have better match scores. They then run the matching algorithm on trajectories from the images of three walking people (9 points tracked per person) where two of the sequences are the same person at different times. They use a subset of points from head, hand and heel to compare match scores. The two sequences of the same person produce higher match scores.

Polana and Nelson [35] present a textural method for detecting periodicity in an XYT solid. Their method assumes a stationary camera. They take the Fourier transform along the time axis and then sum the peak frequencies over x and y to develop an overall periodicity measure for the sequence which they call "activity". They ran tests on 21 sequences that included periodicity (at least four cycles), and 8 that did not. All 21 periodic sequences had higher "activity" than any of the non-periodic sequences. This provides a classification of sequences according to overall temporal periodicity.

Allman and Dyer [3] call their technique "Dynamic Perceptual Organization." It consists of treating an image sequence as an XYT solid, finding the optic flow in the solid; then tracing curves as a function of time, and k-means clustering curves with similar slope and curvature. Data is shown for two test sequences of rotations and translations of textured objects against a stationary background. They then suggest how curves might be decomposed into common and relative motion, and how this might be used for recognition. They show flow curves derived from tracking wingtips of a flapping bird and describe how the curves make apparent the cyclic nature of the flapping. Though the data is suggestive, they do not demonstrate motion identification.

Goddard [17] built a connectionist network to recognize three different gaits. The original signal comes from a WATSMART tracking system, however it is converted to 2D unlabeled spots. The connectionist network takes as input an "augmented optical flow" which includes spot location, velocity, and instantaneous curvature of a path and a "tempo" variable that adjusts the time scale. It has intermediate levels to connect the dots into components and the components into assemblies. The network can identify walking, running, and skipping gaits under a variety of conditions of noise and clutter.

Goddard's work isn't quite model free, but his lack of discussion of any intermediate component structures implies that they were not useful as models, hence we classified it with the model-free techniques. Evaluating it against other recognition systems, we note that three is a small number of categories, and that the strength of the connectionist approach

shows up in its robustness in the face of clutter, but not in the number of motions it identified.

The model-free techniques of Allman and Dyer, and Polana and Nelson are interesting. It is plausible that motion recognition systems could work without any assumptions about objects, at least when there is a small number of recognition categories. A shortcoming of Allman and Dyer's technique is that it requires high quality optical flow, and this is not a solved problem. However, model-free motion can be used as another cue for grouping features into objects, their work suggests ways to do that grouping independent of prior guesses about object grouping.

3.3 Human Body Pose Recovery

The problem of human body pose recovery from real imagery is an important one in computer vision. Some representative work will be discussed here, with an intent to demonstrate that it is as yet too difficult to recognize many different motions in real imagery.

O'Rourke and Badler [34] used constraint propagation to implement a feedback loop between high-level and low-level processes of image motion analysis. The idea is that a high level model of a human body can be used to indicate where in the image to track low level features; the low level features then determine the new state of the high level model. The constraints are used both to limit search of the image and to reject data from low level tracking that is inconsistent with the high level model. Constraints on the high level model include acceleration limits, joint angle limits, and constancy of limb length. Only the hand, foot, and head segments are explicitly searched for in the images. Tests on two sequences are reported; one involving a jumping-jack motion and the other a hand going behind the back and being occluded. In both tests the images are synthetic human figures against a dark background. The system has a complex architecture which these tests did not fully plumb.

Hogg [25] built a system to find and track people walking in real images against complex backgrounds. It contains a human body model made out of cylinders [8]. Walking is modeled as a set of postures (keyframes) where each posture is subject to its own set of positional and velocity constraints. The body model is projected onto an edge map made from the image and "plausibilities" are calculated for each part of the model, and some kind of search

is done to maximize the plausibility of the model. The walking model described used four postures of fronto-parallel walking. it was tested on one sequence of 45 images of a walking person against a complex stationary background, and it found and tracked the person. It is hard to determine, from the single test reported, the degree to which the approach will generalize.

Rohr [38] built a similar system to find and track pedestrians crossing streets in real images against natural backgrounds. It assumes a stationary camera and uses change detection (background subtraction) to fit a rectangle around the walking person. An edge map is made, and line segments are fitted to groups of edge points. The human body model, similar to Hogg's, is used to guide a search of the line segments. The highly detailed walking model was derived by averaging motion curves from medical studies of 60 men, and yielded a walking motion that has just one degree of freedom — pose (canonical time or phase), which varies from zero to one. Information from the walking model and the human body model is used to project windows onto the image, and the window areas are searched for line segments corresponding to the appropriate part of the model. After an initialization search, a Kalman filter is used to estimate three variables: the walker's pose and X and Y position, and these constrain later search. Experimental data is presented for three image sequences: one synthetic and two real, and all involve fronto-parallel walking. The system worked well on these cases. The greatest difficulty was in determining which leg was nearer the camera; it was prone to a pose error of .5.

The work of Akita [2] is similar to that of Hogg and Rohr except that it uses a sequence of keyframe stick figures plus a table describing occlusions for each keyframe, which were entered manually.

The contribution of O'Rourke and Badler is in their sophisticated architecture, which allows bidirectional communication between the high level and low level parts of the system. Each of the remaining systems systems attempts to find humans in real imagery. Hogg and Rohr each had to make a number of assumptions to limit search time. They each assume a fronto-parallel direction, uniform walking gait, and stationary camera. They each capitalize on the stationary camera which allows them to use change detection to draw attention to the rough location and size of the walker.

The problem of seeing people in natural scenes is very hard and while there is much merit in the work of Hogg, Rohr, and Akita it is not clear how to generalize from their systems to

our problem of dance step recognition. One can claim that employing a number of systems in parallel will allow a number of gaits and directions to be searched for simultaneously. However, given hundreds of dance steps and a reasonable number of orientations, this approach would require thousands of specialized recognizers.

3.4 Representation of trajectories

Gould and Shah [18] present a set of representations called the "Trajectory Primal Sketch." Their system takes as input XY data over time of tracked points, and returns as output the identification of various features on the trajectories such as changes in direction or speed. They present what they call the $\Phi - S$ representation which has axes of distance along the curve and local curvature, making the representation independent of x,y position and rotational orientation. However, they do their feature finding on curves of x-velocity and y-velocity (equivalent to $\frac{dx}{dt}$ and $\frac{dy}{dt}$) which are independent of position but dependent on x-y axis orientation. They show how primitive trajectories (e.g. translation, rotation, and cycloids) can be found. They plot trajectories in scale-space, search for significant features (ones which appear at coarse scales), and then annotate the original x,y trajectories with symbols at the points where the features occurred.

The work of Gould and Shah is closer in spirit to the work in this thesis. Their goal is to label features on a trajectory, while our goal is to identify the trajectory. An XY trajectory is a projection of phase space, and their position and angle independent representations are interesting.

3.5 Phase Space

One of the classical techniques for analyzing the dynamics of a system is the phase portrait in phase space. Phase space is a Euclidean space with axes for all the independent variables of a system and their time derivatives. For a given set of initial conditions, the system can be modeled as the motion of a point along a path in this space. Each point in phase space represents a state of the system, and as the system evolves over time it moves along a phase path. For a large set of systems [2] only one phase path passes through each point in the

[2] *autonomous* systems: systems in which parameters such as spring constants do not change with time (if they do, they should be dimensions of the phase space, not constant parameters)

space. This property is particularly useful in systems of a single variable, e.g. a mass on a spring. Such systems have a 2D phase space of position and velocity and the property requires that phase paths can never intersect. So plotting a few phase paths with different initial conditions will often carve up the phase plane and show the behavior of the system for all initial conditions. The family of phase curves is the phase portrait or phase diagram of the system. Sometimes in higher dimensional phase spaces, all the phase paths will lie on a lower dimensional manifold in phase space. Some paths may converge on fixed points which correspond to unchanging states. Some may form closed loops known as orbits, which correspond to periodic motion. For more on all this see Jackson, Chapter 2 [27].

Phase space has been used in the Artificial Intelligence research community in systems which understand and reason about physical dynamics. For example see, The Dynamicist's Workbench [1]. Sacks [39] wrote software that analyzes the qualitative behavior of non-linear ordinary differential equations by emulating the strategies of human experts reasoning about portraits in phase space. Phase space has been used in the animation community for systems which provide more flexible animations [23]. In the computer vision community, phase space has not been widely used as a representation. It was used by Verri, et al. [47] in a theoretical study of properties of optical flow, and, as noted below it has been used by Shavit as a representation of overall motion.

In the computer vision community, Shavit [41, 42] developed a different way to represent qualitative "visual" dynamics using phase space. He considered the binary image of a moving creature to be a 2D picture printed on a blob of elastic material undergoing deformations and rotations over time. From the changes in the image he computed the forces, strains, and rotations that best accounted for the shape changes of the image. The forces give rise to a flow velocity and position that varies with time. He made phase diagrams of flow velocity versus position, and segmented them into sections with simple dynamics (for example, mass on a spring, or viscous damping). Thus, as the system orbited around the phase curve over time, he could identify different sections as having particular dynamics such as linear spring or applied force or damped oscillation. Using inputs such as black silhouettes of walking, running, and jumping cartoon characters; moving light displays of actors performing the same gaits; and some images derived from films of running animals; he was able to identify characteristics about the gaits and to distinguish among the three gaits.

Sclaroff's [40] Modal Matching technique uses finite element analysis to find the vibrational modes of binary images of objects. When it tracks a deforming object it records how much deformation comes from which modes. It then measures similarity as a sum of strain energies over the modes. Given two extremal views (X-rays) of a beating heart, Sclaroff shows a phase diagram of similarity with one extreme view versus the other as the heart goes through the diastole / systole cycle.

Haken, et al. [22] report an interesting psychophysical experiment that had test subjects in a forced choice situation classifying according to relative phase of two angles. Subjects were shown an image similar to an arm which had the elbow and wrist angles in motion. They were shown completely in-phase and out-of-phase motions as target categories A and B. They were asked to classify motions with various phases into either A or B. Haken, et al. report that "Labeling Probabilities changed abruptly along the continuum for all 10 subjects tested. Around a phase difference of 90° the identification function possessed a rather steep slope." This result demonstrates that people are quite sensitive to phase relationships when they look at motion.

The work cited above suggests that phase paths may be a useful representation of motion. Shavit shows how useful information can be extracted from phase diagrams of variables derived from the overall deformation of an image. We will build on this work and show how phase space representations can be useful for identification.

3.6 Other Methods

Other possible methods which could be applied to the problem of dance step recognition include Hidden Markov Models and various techniques from statistical pattern recognition.

Hidden Markov Models (HMM's) are widely used in speech recognition, where they provide the benefits of time scale invariance and a way of combining segmentation and classification into one task. HMM's were used by Yamato et al. [52] to recognize tennis strokes within a vocabulary of six strokes where the input was a subsampled binary image. This means their representation of a tennis stroke is a sequence of images with a particular camera angle, background, clothing, and lighting. HMM's operate on sequences of symbols, so some kind of feature must be extracted from the signal and fed to the HMM learning algorithm. In the work of Yamato et al. features were grey level changes, making their

recognition entirely dependent on details of lighting, etc. so it is not clear if their system is classifying movements or if it is only classifying views. If HMM's were applied for the tracking data used in this thesis, input features could include peaks and zero crossings of velocities and accelerations of joint angles.

A major advantage of HMM's is that the signal does not need to be segmented before recognition. A major disadvantage is that many training instances must be supplied. Yamato et al. used 30 instances of each stroke (10 from each of 3 players; 5 for training, 5 for test). In their test they presented their HMM's with separate data sequences for each stroke. By eliminating the segmentation task during recognition, (as well as eliminating any context dependent variations in the motion), the HMM needs less training. As the vocabulary and difficulty of segmentation increases, the number of training instances must also increase. Starner et al. [43] used HMM's to recognize cursive handwriting. Their training set included 25,600 words in cursive. The necessity of such large training sets make HMM's inappropriate for this project.

A typical pattern classification technique such as Maximum Likelihood [14] (or Baysian classification if unequal priors are introduced) would begin by forming an n-dimensional vector from the parameters for each dance step. The training instances for a given step would map to a cluster of points in n-space. The centroid of that cluster would be the canonical instance of the step, and its covariance matrix would be used to compute the likelihood that a new point is a member of the cluster. Once all steps were learned in this way, a new step could be classified by calculating likelihoods from all learned steps and picking the maximum. The difficulty of applying this technique to ballet data is that the stream of data must be segmented into separate dance steps before it can be formed into a vector and classified. All the difficulty in this approach becomes concentrated in the segmentation phase.

3.7 Summary

Little work has been done in the area of recognizing whole body movements. The work that has been done involves only a small number of categories. Rohr, Hogg, and Akita assume a single type of motion (walking) and search for it, while Goddard and Shavit each assume three categories. This small a number of categories does not adequately test the ability of

the representation or the recognition system to discriminate between different movements. Yamato uses six categories, but he may be classifying views rather than movements.

Other work that has been done focuses on developing an intermediate representation of movements. Webb and Aggarwal, and Hoffman and Flinchbaugh can be seen as extracting an axis and angle from a set of coordinates of an object; Gould and Shah, and Shavit segment trajectories and characterize the segments or their endpoints. However, only Shavit uses his representation for recognition, and the others do not use theirs for any further task and so cannot be evaluated.

The task of this thesis involves more recognition categories, and this puts greater demands on the representation. Phase space, which was used by Shavit, has many valuable features (which will be discussed below) and a representation based on it is used in this thesis. However Shavit's model of an object as a deforming rubber slab is not adequate for ballet and so a more complex 10 joint, 24 DOF human body model is used in this thesis.

Chapter 4

Representation and Recognition in Phase Space

A recognition task can be rendered simple or complex depending on the choice of representation, and a representation can be judged on how it facilitates the task to which it is applied. In this Chapter our representation of human body motion is presented, followed by the selection criteria and desiderata.

Consider several occurrences of a movement; e.g. a plié, in which the legs repeat exactly the same motion, while the arms behave differently each time; and consider plotting all the movements in the same full phase space of all the torso position and attitude parameters and joint angles, and all their derivatives, and no explicit time axis. Plotting only the leg variables while holding the arm variables constant would show a space curve, and each repetition of the movement would traverse that same space curve. However, if the arm variables are plotted as well, each repetition may traverse a different space curve. The intuition here is that the invariance of the plié is captured in the leg variables, while the variation is in the arm variables. Equivalently, the invariance can be found in a subspace of the full phase space. This intuition motivates our choice of representation for movement. Figures 4-1, 4-2, and 4-3 show 2D projections of space curves in phase space for three ballet movements embedded in the phase plots of other motions; using joint angles extracted from XYZ data (see Appendix A for details).

We represent movement by a space curve in a subspace of phase space. The subspace is obtained by a trivial projection parallel to certain axes (such a projection is equivalent

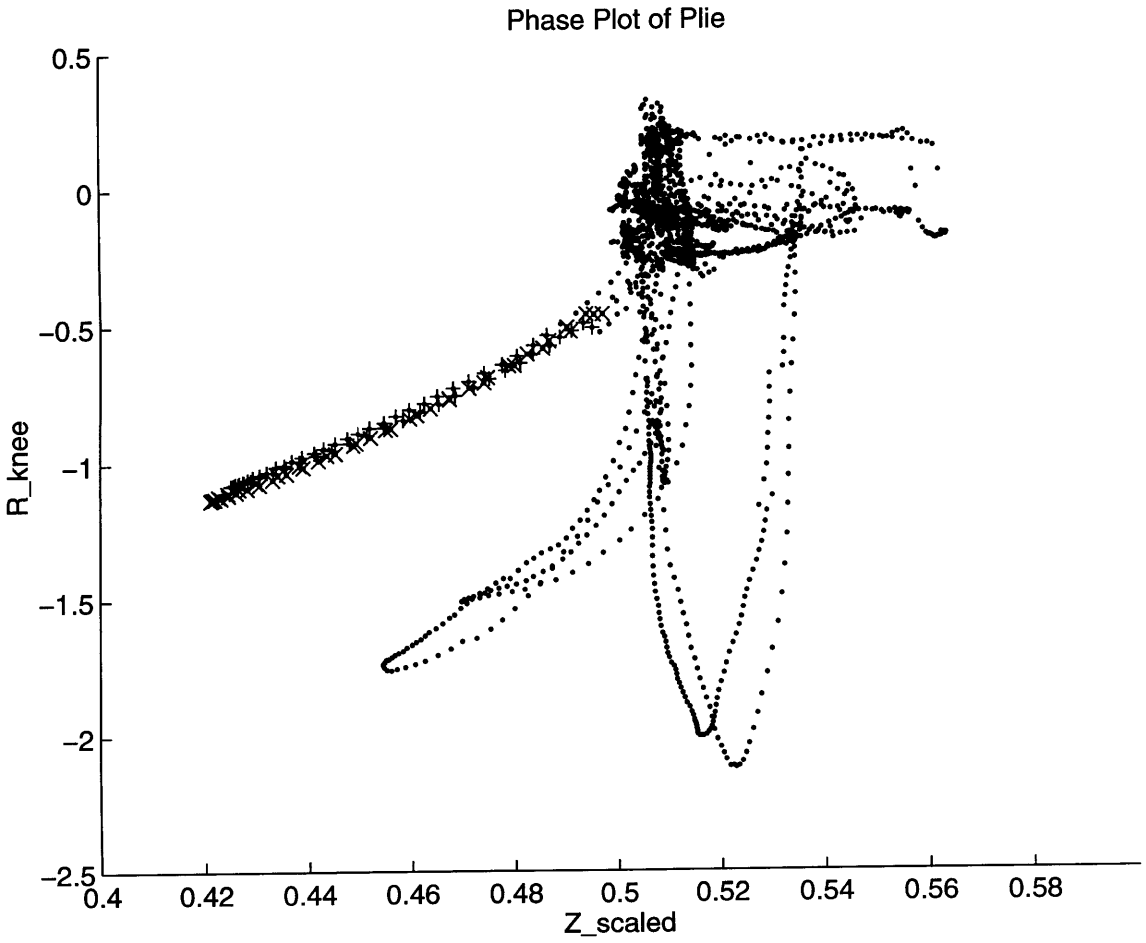

Figure 4-1: $\times$ and $+$ marks time steps during pliés for two dancers; "." marks time steps during other movements.

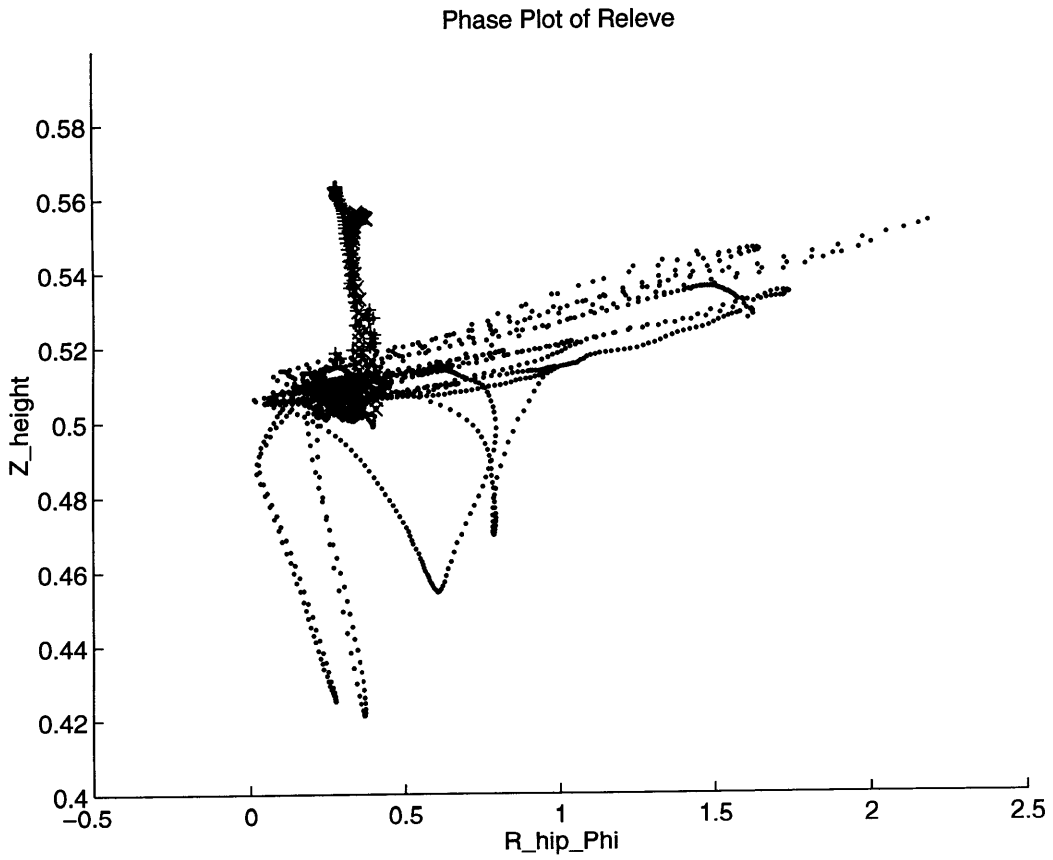

Figure 4-2: $\times$ and $+$ mark time steps during relevés for two dancers; "." marks time steps during other movements.

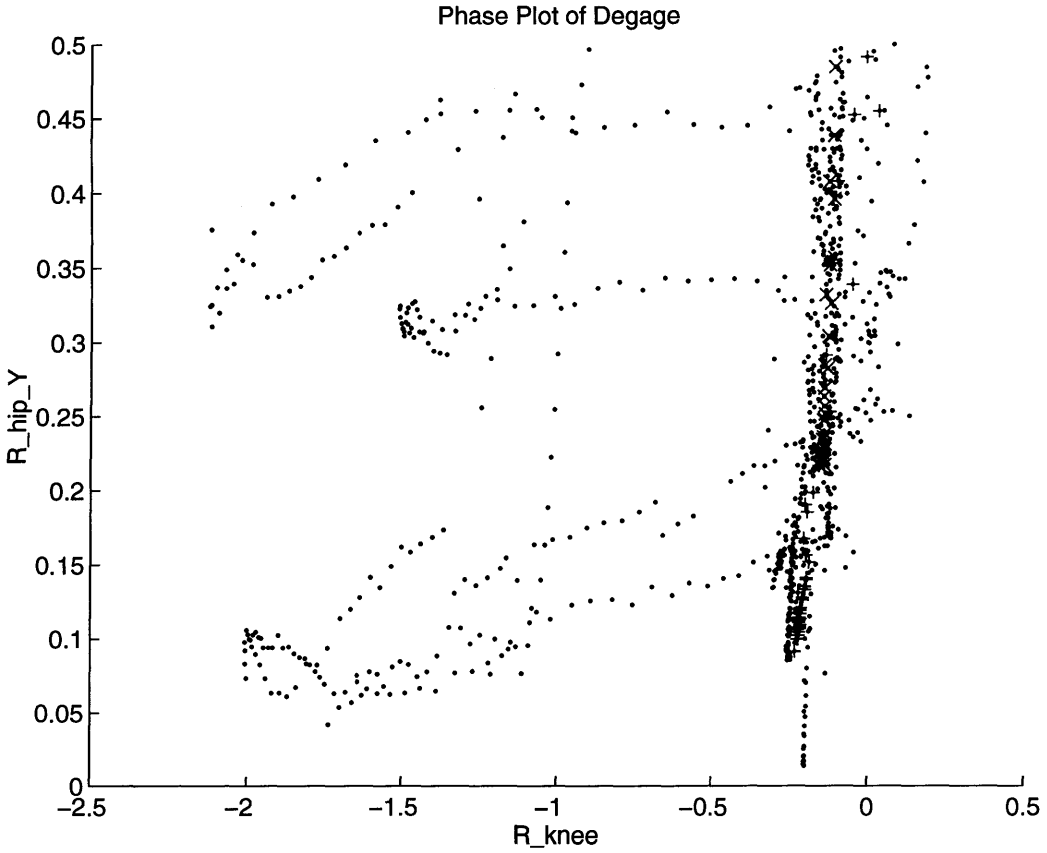

Figure 4-3:
$\times$ and $+$ mark time steps during dégagés for two dancers; "." marks time steps during other movements. Dégagé sweeps the leg out to the side and back to vertical; tendu, développé, and grand battement all end with a similar sweep of the leg back to vertical, though the sweeps start from different heights.

to ignoring the coordinates along those axes). Our method of selecting the subspace is described in Section 4.4.

We parameterize the space curve as a smooth and continuous curve. In the current implementation the parameterization is an intersection of manifolds, each of which when projected onto the appropriate 2D plane makes a segment of a cubic curve of two variables. However, the representation admits any continuous curve parameterization such as piecewise linear segments or b-splines.

Recognition occurs for a sequence of time steps. At each time step the state of the system can be represented as a point in phase space. We conditionally accept a point as being part of a recognized movement if it is within a threshold distance from a space curve in the projected space (actually there is a different threshold for each axis of deviation, so the accepting volume is swept out by an n-dimensional rectangle). We accept a sequence of time steps as being a recognized movement if all the points in the sequence stay within an accepting volume for more than a duration threshold. In the present implementation, velocity is not considered when a sequence is accepted.

The 1978 publication of Marr and Nishihara contains valuable insights for selecting a representation for vision, and the 1979 publication of Badler and Smoliar contains valuable insights on representation of movement. We shall evaluate the general form of our representation according to the sets of criteria published in these two works.

4.1 Representation Considerations

Marr and Nishihara [31] discuss representing three dimensional shapes for object recognition, and they suggest three criteria and make three recommendations. Badler and Smoliar [5] discuss representations of human movement for purposes of animation, simulation, and translation. Though their purposes are slightly different from ours, the considerations they put forth proved valuable in this thesis, and so we will evaluate our representation in light of their criteria.

Marr and Nishihara present three criteria: *accessibility*, *scope and uniqueness*, and *stability and sensitivity*; and their three recommendations are: use an *object centered coordinate system*, include volumetric *primitives of varied sizes*, and have a *modular organization* [emphasis *mine*].

- *Accessibility*: is it easy to compute the representation from the raw data?

Yes, joint angles are easily computed from tracking data, and our learning algorithm easily finds regions of phase space.

- *Scope and uniqueness*: what class of movements is the representation designed for, and do the movements have canonical descriptions in the representation?

For an autonomous system, a space of torso position and attitude parameters and joint angles and velocities will describe all possible motions uniquely. This is equivalent to the statement that only one phase path passes through each point in the space. A dancer exerting forces on the ground is not an autonomous system, so multiple phase paths can pass though the same point, but if two whole paths are coincident, they are considered the same motion.

However, since two different dancers will probably not traverse identical phase paths, a question still remains: can two motions judged to be the same by a human observer also be placed in the same category by a system using this representation? In other words, is there a way to represent the commonality between movements, as well as the differences? The ability to project out (ignore) dimensions of phase space which do not help prediction helps to identify commonality, but this is only a partial answer. As will be seen in Section 4.4, for each movement to be learned, we ignore velocities and search all possible pairs of a set of configurational parameters, $N(N-1)$ 2D subspaces, to accumulate a set of parameters which describe the "invariant subspace" for that movement.

- *Stability and sensitivity*: Does degree of similarity in the representation reflect similarity in the motions? And can subtle differences be expressed in the representation?

The representation is capable of expressing degrees of similarity by some metric of distance between two paths in phase space. However, the Euler angle representation of 3DOF joints such as hips and shoulders has singularities at certain orientations, and near these singularities the sensitivity becomes unacceptably high (near singularity, an arbitrarily small change in the tip of a vector can cause two of the Euler angles in the representation to make maximal 180° swings). One solution to this sensitivity problem is to use multiple different sets of Euler angles to represent 3DOF joints, and to do recognition in the angle set which is not near singularity. Nearness to singularity can always be detected because

the axis of the third Euler angle becomes nearly parallel to the axis of the first Euler angle. If θ is the angle between the first and third axes, either $1/\sin(\theta)$ or $1/\cos(\theta)$ is a measure of sensitivity, and an indicator of when to switch representations. This will be discussed in greater detail in Section refreporient.

- *Object centered coordinate system*:

The hierarchical human body model we use is object-centered.

- *Primitives of varied sizes*:

No, this is a weakness of our system. Our only primitive is a single segment of a cubic curve. This is adequate for "atomic" motions but not for more elaborate compound motions.

- *Modular organization*:

The hierarchical human body model is modular. The idea of atomic motions is modular, but not all dance steps are built out of atomic motions. Thus we have partial modularity.

Badler and Smoliar present four considerations relevant to recognition:

1. Both destinations (goals) and movements (changes) can be specified.
2. Movement can be constrained by described relationships between body parts or other objects such as physical contact, proximity, and surrounding.
3. The dynamics and phrasing of a movement should be separable from the spatial displacement of each body part.
4. The system should be tested on movement sequences noted for their scope and variety.

1. *destinations and movements*: Our representation cannot describe destinations without movements; however this criterion of Badler's is directed at ergonomic studies where the purpose of a motion is paramount. In ballet, destination and path are equally important.

2. *constrained by described relationships*: our representation provides such constraints only in the human body model which enforces connection of limbs and prevents elbows and knees and ankles from bending beyond straight.

3. *dynamics and phrasing separable*: Our representation makes this explicit since phase space has separate position and velocity axes.

4. *tested on sequences of scope and variety*: this is good advice in general, however we are testing on a relatively small number of dance steps. It would be easy to prepare a widely disparate test sequence which would make recognition easier, but a better test of the system's ability to make fine distinctions would include some steps that are close together.

One question still remains: why use a phase space of joint angles as opposed to some other set of variables? For the recognition method to work, we need successive repetitions of a motion by different dancers to lie close to the same space curve in phase space, and thus need to represent or describe motion using a set of variables which will give us this property. Such a description needs to be independent of position and orientation of the torso, and it is certainly useful if the description of one limb is independent of the descriptions of the other limbs. Joint angles are a simple representation that has these properties.

4.2 Recognition in Phase Space

In order to represent motions in phase space, we must first convert the XYZ tracking data to joint angles and torso position and attitude parameters. The mechanical details of collecting tracking data and converting to joint angles are described in Appendix A. XYZ data for 14 points is recorded, and joint angles for six 1DOF and four 3DOF joints in the arms and legs are computed.

Any body motion can be represented as a space curve in the full phase space [1]. However, repetitions of the motion will not trace the same space curve unless they are performed at the same location, attitude, and speed. The representation can be made invariant with respect to location and attitude by projecting out those six variables and their derivatives which correspond to torso position and attitude. The representation can be made invariant to speed of performance by projecting out all time derivatives. Similarly, if arm positions do not repeat during different repetitions of a body motion, but leg positions do, then all variables corresponding to arm positions can be projected out, leaving a much smaller phase space. Note that projecting out time derivatives does not remove all considerations of dynamics and velocity: relative velocities must be preserved if relative angles are invariant

[1] the full phase space is composed of 48 dimensions, the following 24 parameters and their time derivatives: torso location and attitude: 6 parameters represented by X, Y, Z, yaw, pitch, and roll; shoulders and hips – 3 parameters each, represented by 3 Euler angles; elbows, knees and ankles: 1 parameter each represented by 1 angle. See Figure 5-1.

with respect to velocities.

The smallest usable phase space is a 2D phase plot, and that is where this method begins. Some of the advantages of a phase plot with the above mentioned variables projected out are:

- These plots will be independent of the speed of performance, because speed does not affect the phase relationship of the angles;
- They will also be relatively independent of the degree of extension with which a movement is performed if the right pair of variables is chosen;
- There is no need to segment the movements before attempting to identify as long as each movement has its "own" region of phase space.

The phase plot will only work perfectly when there is one degree of freedom in the system; in other words, when it has an intrinsic dimensionality of 1. For example, in a perfect plié, the center of gravity is raised and lowered along a vertical line passing midway between the feet. The hip, knee, and ankle joint axes of both legs are parallel, so all are constrained to move through the same sequence of configurations each time the motion is performed.

Though most real world moves are not constrained to one degree of freedom, the constraints of balance and support frequently combine to produce a "most natural" or "most comfortable" motion with one degree of freedom for the major body masses. Furthermore, in ballet the correct forms of many movements require that they be performed as if constrained to one degree of freedom. For example, dancers have the option to move their center of gravity from side to side (as long as it remains above their feet) but ballet posture requires that it be centered. When a leg is extended for a weight transfer, the toe is brushed along the floor (tendú) while the ankle bends to maintain toe contact. when a leg is raised for a pirouette, the instep of the raised leg slides along the supporting leg to rest near the knee. Each of these moves has just one degree of freedom, and there are many more such examples.

Sometimes only subsections of the body move with 1 DOF: e.g. in a plié the arms are unconstrained but the legs and torso move with 1 DOF. So the problem is to find those variables which encode the 1 DOF. This is done by using training data with dance steps

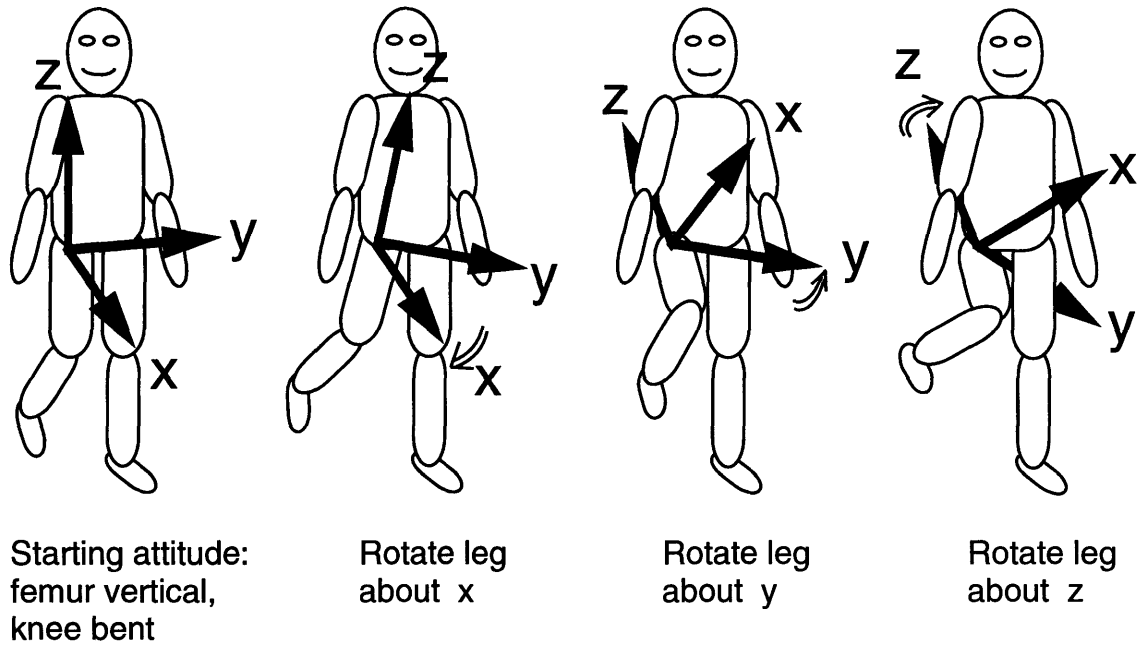

Figure 4-4: Three successive rotations illustrate Euler angles representing 3 rotational degrees of freedom of hip joint. Coordinate axes are rigidly attached to femur, i.e. rotations are measured in frame of reference of leg.

already identified, searching all the $N(N-1)$ potential 2D phase plots to find the ones that make the best predictors, and saving a list of the best predictors. This procedure is explained fully in Section 4.4 on learning.

The recognition system we implemented runs an independent detector for each dance step. Thus it is not a pattern recognition system because it is not prevented from classifying the the same input data as two different dance steps.

4.3 Representation of Orientation

Hip and shoulder joints have three rotational degrees of freedom, and there are several methods available to represent the orientation of objects with three rotational degrees of freedom. The most widely used include rotation matrices, Euler angles, and quaternions (also known as the Euler parameters or Euler's symmetric parameters representation) [26, 30]. In each of these representations, the values of the parameters can be thought of as measuring an angular displacement from an initial orientation. The problem of representing orientation occurs in disciplines such as aircraft and spacecraft guidance, robotic control, and computer graphics, and in these different domains different choices of representation

are often made.

A rotation matrix $\mathbf{R}$ is an orthonormal 3×3 real matrix and thus has nine parameters while only three degrees of freedom. It can be thought of as a composition of three successive rotation matrices representing rotations about three orthogonal axes, e.g.:

$$\mathbf{R}_x\mathbf{R}_y\mathbf{R}_z = \mathbf{R}.$$

where the initial orientation is represented by a unit matrix. The components are products of trigonometric functions of the rotation angles; in the above example

$$r_{33} = \cos\theta_1 \cos\theta_2 \cos\theta_3 - \sin\theta_1 \sin\theta_3.$$

Rotation matrices are common in computer graphics but they are a poor choice of representation for recognition because the three degrees of freedom are distributed over nine parameters, so all nine parameters must be considered together to recognize an orientation. Since there are only three degrees of freedom, it is more straightforward to consider a three parameter representation.

Euler angles $(\theta_1, \theta_2, \theta_3)$ consist of three parameters representing angles of rotation about the three coordinate axes (see Figure 4-4). For small deviations from the nominal they are independent, and for a significant range of orientations they are essentially "linear" e.g. similar sized movements of the knee in that range result in similar sized changes in the Euler angle values. There are 12 possible sets of Euler angles: six are similar to the axes shown in Figure 4-4 where all axes are perpendicular when $\theta_1 = \theta_2 = \theta_3 = 0°$; they come from the six possible permutations of the X, Y, and Z orthogonal axes and may be denoted X–Y–Z, X–Z–Y, Y–X–Z, Y–Z–X, Z–X–Y and Z–Y–X. The other six from "polar coordinate" representations where when $\theta_2 = 0°$, θ_1 and θ_3 are parallel, and may be denoted X–Y–X, X–Z–X, Y–X–Y, Y–Z–Y, Z–X–Z, and Z–Y–Z. Euler angle representations all suffer from a problem of singularities: there always exists a value of θ_2 such that θ_1 and θ_3 are parallel. When this happens there is no longer a canonical representation for the orientation – rotations about the two parallel axes can be allocated arbitrarily to θ_1 and θ_3. A worse problem is that near singularity, arbitrarily small changes in orientation can cause maximal changes in θ_1 and θ_3. This means that the representation has too much sensitivity to noise near singularity (singularity occurs at $\theta_2 = 90°$ for the initially perpendicular sets, and

$\theta_2 = 0°$ for the "polar coordinate" sets).

Quaternions are based on Euler's theorem, which states that any angular displacement, which can be composed of arbitrary rotations about arbitrary axes, can be expressed as a single rotation about a particular axis. Quaternions consist of four parameters (n_1, n_2, n_3, p), three of which represent an axis, while the last represents rotation about that axis. There is a constraint that $n_1^2 + n_2^2 + n_2^2 + p^2 = 1$. The values of the parameters can be found thus: let $\hat{\mathbf{v}}$ be a unit vector parallel to the axis; let ω_1, ω_2, and ω_3 be the projections of $\hat{\mathbf{v}}$ on the coordinate axes (the so-called direction cosines); and let ϕ be the angle of rotation about $\hat{\mathbf{v}}$. Then

$$(n_1, n_2, n_3, p) = (\omega_1 \sin\frac{\phi}{2}, \omega_2 \sin\frac{\phi}{2}, \omega_3 \sin\frac{\phi}{2}, \cos\frac{\phi}{2}).$$

Quaternions have no singularities, but they do have some curious properties: there are two representations for every orientation (negating n_1, n_2, n_3, and p yields a different quaternion but the same orientation), and as the angle approaches zero, the axis parameters become small and lose sensitivity.

Many ballet movements involve rotating the hip joints about a single axis axis such as the left-right or front-back axis of the torso. These axes can be aligned with an Euler angle coordinate system fixed to the frame of reference of the torso so that the movements cause changes in a single Euler angle. Thus if the problem of singularities can be solved, Euler angles make a good choice of representation for ballet. As is noted above, one can use multiple different sets of Euler angles to represent 3DOF joints, and do recognition in the angle set which is not near singularity. The second Euler angle can be used to detect when to change representations: $1/\sin(\theta)$ indicates sensitivity for perpendicular axis representations and $1/\cos(\theta)$ for polar coordinate axis representations, so, e.g. changeover from a perpendicular axis representation to a polar coordinate axis representation can occur as θ_2 becomes greater than 45°.

An additional advantage of Euler angles for representing ballet data is related to missing data. When the ankle marker is obscured but hip and knee markers are visible, only partial data can be recovered for hip orientation (at most two out of the three Euler angles depending on the representation). For Euler angles the partial data can be computed and used for recognition, but for quaternions no values can be computed from the partial data. Thus Euler angles are slightly more robust in the presence of missing data. Quaternions

are interesting because they have no singularities, but their use in recognition remains for future work.

4.4 Learning

The learning phase of the system is supervised learning in which the system processes input where each time step is previously identified as being part of a particular dance step, or as part of a non-dancing pose. To find good predictors, the system does a hierarchical search, so that it can bound a region of phase space as the accepting region of a movement. There are several intuitions behind the algorithms used, having to do with the nature of correlations, a way to limit search, and a way to find successively smaller regions of phase space.

The algorithm operates on a list of parameters including joint angles and torso position and attitude parameters. Some parameters, (e.g. torso position on stage) are removed from consideration using a priori knowledge that they are not part of the "invariant subspace."

Recall that for each time step there is a point in the full phase space which can be projected into a lower dimensional phase space, and that there are instances of each movement in the ground truth training data. With this in mind, some terms can be defined:

Pair Relation: f_{ij} a smooth function $\widehat{\psi_j} = f_{ij}(\psi_i)$ where ψ_i is an input parameter and $\widehat{\psi_j}$ is a predicted parameter.

Threshold: a distance h above and below a pair predictor $|\psi_j - f_{ij}(\psi_i)| \leq h$ which bounds the accepting region of that predictor.

Pair Predictor: a binary function of time $g_{ij}(t) = \begin{cases} 1 & \text{if} \quad |\psi_j - f_{ij}(\psi_i)| \leq h \\ 0 & \text{otherwise} \end{cases}$ composed of a simple relation and threshold.

Pair Prediction: the output of a pair predictor.

Smoothed Predictor: a filter $g^s_{ij}(t)$ over the pair prediction which eliminates short periods of acceptance or rejection which are less than a set time constant.

Predictor Fitness: a function $\mathcal{F}(g^s_{ij}(t))$ used to evaluate pair predictors and to order them. Predictor functions used in this thesis take the form $\mathcal{F}(g) = wA_{\mathrm{f}}(g) + R_{\mathrm{f}}(g)$ where

$A_f(g)$ and $R_f(g)$ are the number of false acceptances and false rejections respectively, and w is a weight factor.

Compound Predictor: a function $g^*(t) = g^s_{ij}(t) \wedge g^s_{pq}(t)$ where $i \neq j$, $p \neq q$, and $\{i,j\} \neq \{p,q\}$; created by logically anding together the smoothed predictions of two or more pair predictors.

For each movement to be learned, the system considers all possible pair predictors for that movement, finds the best threshold for each pair predictor (evaluating predictors according to a fitness function), and saves the k best pair predictors. Several of these are then combined to form a compound predictor for the move. Central to this process is the fitness function, which must evaluate the quality of correlation revealed by a predictor.

4.4.1 Correlations and Fitness Functions

Correlations in the world lead to categorizations. Correlations come in two types: always true, e.g. laws of physics or math, and true when an item is a member of a category – useful for classifying [9]. The learning method tries to find correlations between variables which occur during a move. But how to find the second kind of correlations which are useful for categorization? Our technique is to minimize a weighted sum of false acceptances and false rejections of a predictor. The false rejection part of the fitness function seeks a good correlation during the move. The false acceptance part of the function seeks anti-correlation during the non-move. Thus this rule finds the second kind of correlations and not the first.

The ability to find functions correlating with category membership allows some freedom in choice of representation – we can allow multiple different representations of the same variable. Although two representations of one variable may be perfectly correlated at all times, they make a poor predictor of category membership and so are not selected by the learning rule.

The same fitness function is used for all ballet steps, both to select the best threshold for each pair predictor, and the k best pair predictors from the set of all pairs. These are different uses with different strengths and weaknesses, and will be discussed separately.

The fitness function determines the threshold; the threshold represents a window about a curve fitted to the training data, corresponding to the idea that a certain amount of noise or variation must be tolerated in each parameter measurement. In the current implementation,

the threshold is fixed, independent of position along the curve. This is a shortcoming which fails to acknowledge the fact that variation is a function of Euler angle sensitivity, of the dancer's degree of control in various positions, and of differences in kinematics of dancer's bodies.

The fitness function also determines which pair predictors are the k best. Since there is a finite number $N(N-1)$ of pair predictors and the pairs are searched exhaustively, the best one, as measured by the fitness function, is found. However, different fitness functions serve different goals. We found that compounding eliminates many more false acceptances than correct acceptances, so a lower penalty could be put on false acceptances if two way or three way compounding was used than if no compounding was used. The intuition here is that true acceptances will be be correlated, so they will tend to be accepted after logical anding, but false acceptances will be uncorrelated and so will tend to be eliminated by anding. For example: consider a sequence of 1000 time steps where a plié occurs for 50 time steps. Assume pair predictors A and B each have 90% correct acceptance and 10% false acceptance. If the predictors are completely uncorrelated, then after anding predictions of A and B one would expect 81% correct acceptance and 1% false acceptance. In practice the false acceptances tend to be partially correlated, but anding predictions still decreases the false acceptance rate in most cases. Anding can also be interpreted in terms of phase space volumes, as will be seen below.

The fitness function handles missing data by treating it as rejected by every predictor, but with no score or penalty for correct or false rejection of missing data. Missing data can be accepted by the smoothing filter, and if it is falsely accepted it will be penalized in the same manner as any other false acceptance.

4.4.2 Volumes in Phase Space, Compound Predictors, and Search

The acceptance region of a pair predictor is represented by an area about smooth curve in a 2D projection of the full N dimensional phase space. This means that in the ND phase space the acceptance region sweeps out a tube-like volume which is bounded only in the two dimensions of the pair predictor. When the predictions of two predictors are logically anded together, the acceptance volume of the resulting compound predictor is the intersection of the acceptance volumes of the component predictors. This is a smaller volume and is bounded in more dimensions, which is why compound predictors are more

conservative discriminators.

In the current implementation, the threshold of a pair predictor is a distance from a cubic curve measured parallel to the dependent axis. A predictor compounded from two pairs has two thresholds. Its acceptance region is a curving tube of rectangular cross-section which, when sliced with the plane defined by the two dependent axes, yields a constant rectangle whose half-dimensions are the two thresholds. The fact that the cross-section is rectangular instead of ellipsoidal, and the fact that the cross-section is constant are both limitations of the implementation, not of the method. In the current implementation, compounding is done by logically anding the three best pair predictors, and thresholds are not altered after compounding. These are both shortcomings; it would be preferable to try compounding all of the k best predictors and to search for a local optimum in thresholds of the compound predictor.

Since the result of the learning process is a multidimensional tube-like volume in phase space, the question arises: why construct it by compounding instead of directly optimizing all parameters simultaneously? The answer is that the compounding approach limits search. The fitness measure as a function of threshold does not necessarily have a single minimum (i.e. the receiver operating characteristic is not convex), so a large range of values of each threshold need to be searched. If multiple thresholds were searched simultaneously, it would lead to combinatorial explosion. The approach we use is a sequence of local optimizations which is not necessarily a global optimum, but the global optimum is computationally intractable.

4.5 Summary

There are several benefits to the phase space representation: it can be made invariant to the speed of a movement by projecting out the speed axes, and it is invariant to the extension of a movement. Thus the representation simplifies the recognition task. The representation is easily computed and can be learned automatically from ground truth data.

Learning occurs in three areas: learning a threshold for each pair predictor, learning which pair predictors are best, and learning which compound predictor is best. At present, the first two kinds of learning are implemented by maximizing the fitness function, and the third is implemented by anding the three best pair predictors. There is a fourth area of

learning which is not automated: the experimenter must learn what fitness function to use.

The major shortcoming in the representation comes about because projecting out velocity variables removes dynamics from the representation. This is not a problem when the system is not presented with distractors (a distractor in this sense is a non-ballet motion; e.g. a motion that starts as a plié but halts or oscillates). Some techniques for handling velocity data are considered in Chapter 6.

Chapter 5

Experiments

5.1 Introduction

We have presented a system for representation and recognition of ballet steps. From the design goals of the system come these criteria for evaluating the recognition system:

Discriminability: that it can classify a fairly large number of steps; our goal is 20 and we have tested for nine in the current implementation.

Abstraction: that it can learn ballet steps from several dancers and recognize new instances of the learned steps from those dancers.

Noise Tolerance: that the recognition system is robust to small errors in tracking data;

Speed Invariance: that the system be able recognize movements performed faster and slower;

Extension Invariance: that the system be able recognize movements performed to different extensions;

Segmentation: the degree to which the system can classify each time step into the correct dance step category (or the non-dance category).

In addition to the above criteria, it is convenient for development and testing to have a system which is easy to train. We designed our system to learn new steps from ground truth data automatically.

A test of a recognition system for the above properties involves: training the recognition system on several sequences of the dance steps from several different dancers, and attempting recognition on several other sequences from the same and different dancers. Both the training data and testing data must be annotated with ground truth data indicating when each step begins and ends.

The tests will be scored in two different ways:

- *Correct timing:* measure the time step at which recognition system indicates that each step began and ended, and the time difference from the ground truth beginning and ending times. Deduct these times from the score.

- *correct labeling*: Let the "announce time" be defined as the time halfway between the time when recognition system indicates a ballet step begins and ends. Count a success each time the announce time occurs some time during the ground truth duration of that step. Count a mis-labeling error each time the wrong step is announced during a step. Count a false positive error whenever a step is announced and no step is occurring. Count a double positive error if a step is correctly identified, but announced twice during a single step. subtract 1 from the score for each error.

As noted earlier, the recognition system we implemented runs an independent detector for each dance step; thus it may err by labeling the same input data as two different dance steps. Tests are always scored for correct labeling, and those which are correctly labeled are scored for correct timing.

5.2 Testing

Our tests were resubstitution estimates in which the same data is used for training and for testing.

Input data was obtained from the multitrax tracking system which tracked markers attached to the main joints of ballet dancers (toes, ankles, knees, hips, shoulders, elbows, wrists). Details of the tracking setup are described in appendix A. Our system extracted joint angles from the tracking data according to a human body model with 1DOF joints at ankles, knees, and elbows, and 3DOF joints at hips and shoulders, plus six more coordinates for torso location and attitude.

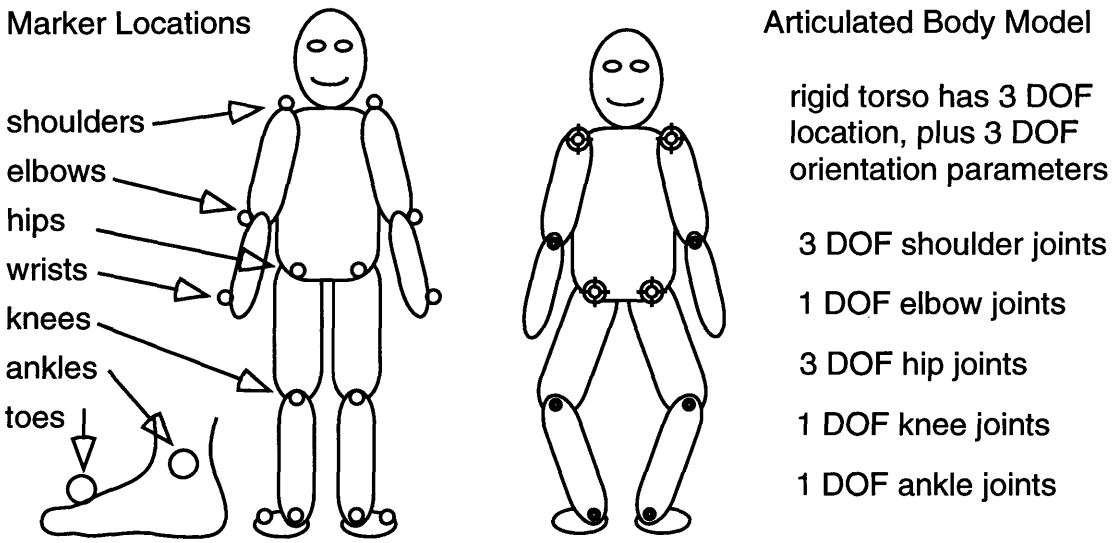

Figure 5-1: Tracking and human body models.

Data was recorded from two different women dancers, of height 157 and 173 centimeters. This is enough of a size difference so they had differently proportioned feet. One was a serious amateur and one a professional ballet dancer.

As detailed in Section 2.3 the ballet steps to be recognized are: plié, relevé, tendu, dégagé, fondu, frappé, développé, grand battement á la seconde, and grand battement devant.

Training data consisting of joint angles and ground truth annotations are sent to the learning system which is described in Section 4.4. Figure 5-2 shows joint angle data for the nine ballet steps for the right hip, knee, and ankle joints. Spikes in the values are missing data codes (missing data receives zero weight in the learning system). The hip-Z angles are quite noisy; this is because when a dancer points her toe and the knee is straight, the hip, knee, ankle, and toe markers are nearly collinear and thus near singularity. Noisy angles make poor predictors, so the learning system does not choose them as part of a predictor.

Learning occurs in three areas: learning a threshold for each pair predictor, learning which pair predictors are "best" (i.e. maximize the fitness function), and learning a compound predictor. At present, the first two kinds of learning are implemented by choosing the threshold or pair predictor that evaluates highest in the fitness function, and the third is implemented by anding the three best pair predictors. There are two parameters which must be set manually: the fitness function weight ratio, and the smoothing time constant.

The fitness function is a negative weighted sum of false rejections and false acceptances,

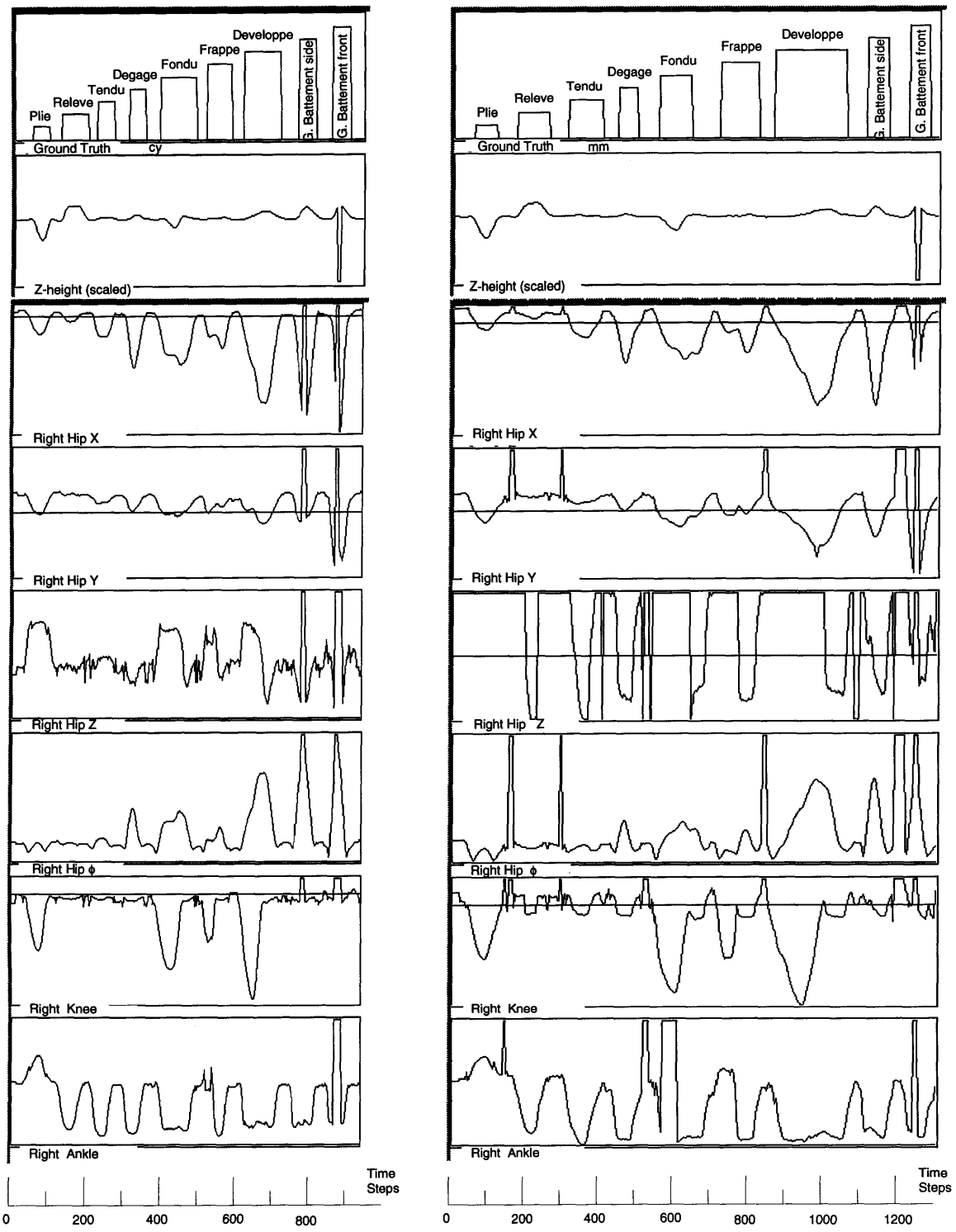

Figure 5-2: Input data for two dancers processed by dance step recognition system. The top graph is ground truth identification; the next is torso height scaled by dancer's leg length; three Euler angles representing the dancer's right hip; a fourth Euler angle from a different representation; and right knee and ankle angles. The time base along the bottom is measured in video frames. The "spikes" clipped at the tops of graphs (e.g. in Right Hip ϕ) are missing data codes.

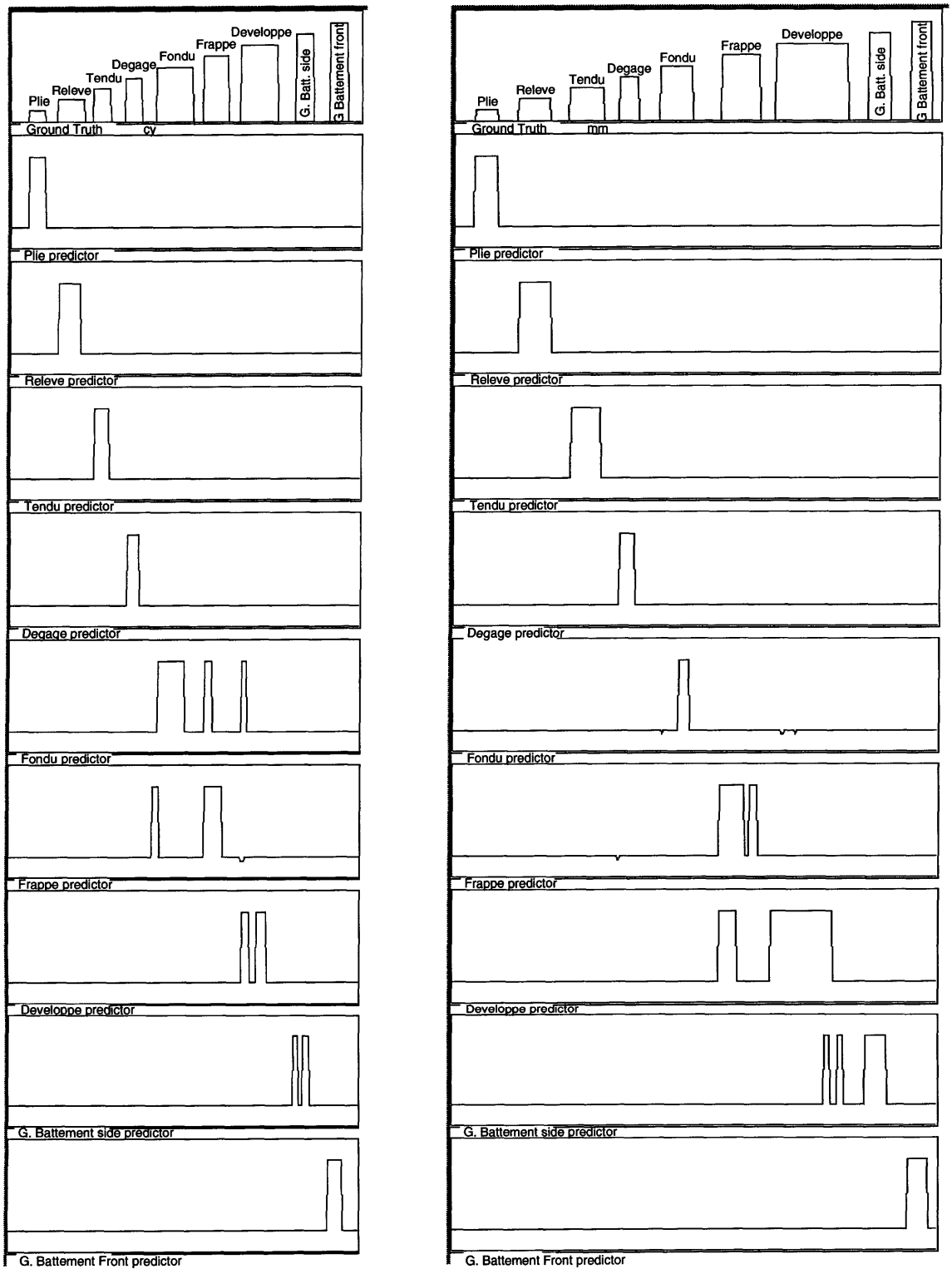

Figure 5-3: Predictor outputs for two sets of movements. Note false positives from predictors 5 and 6 (fondu, frappé) on the left chart, and 7 and 8 (développé, and grand battement side) on the right. Also note the "double announcement" of step 8 on the left chart.

so the degree of freedom lies in choosing the relative weights. For the data we tested the system responded similarly to weight ratios in the range of 4.5:1 to 6.3:1 (false rejection penalty : false acceptance penalty). The final version of the fitness function had a ratio of 5.9:1. Though the relative weights in the fitness function might appear to favor false acceptance it is explained by the effects of compounding (anding) of predictors: compounding can never increase acceptances; they must either decrease or remain unchanged. It was found that compounding decreased false acceptances much more than correct acceptances (see Section 4.4.2 on phase space volumes) so the fitness function could be set to allow many more false acceptances than false rejections with the expectation that most would be rejected during compounding.

A smoothing time constant is used to remove short periods of acceptance or rejection. A time step is not accepted as part of a dance step alone; it must be part of a sequence of accepted time steps, and the sequence must be longer than the smoothing time constant to be accepted. Similarly, a gap in a sequence of accepted points will not be rejected unless it is longer than that same time constant. The value for the smoothing time comes from the idea that ballet steps must have a duration of a sizable fraction of a second, i.e. no ballet movement is performed in less than a tenth of a second. We tried values in the range of .1 to .2 seconds, and used .2 seconds in the final version.

Figure 5-3 shows predictions for all nine ballet steps, and Table 5.1 summarizes the scoring. Steps 1, 2, 3, 4, and 9 are predicted extremely well; the average error for begin and end times is 4 time steps or 67 msec. Predictors for steps 5, 6, and 7 all falsely predicted in addition to their correct predictions. This is because all three are compound movements, and are poorly represented by simple cubic curves. When the curve fits poorly, the system must allow a larger threshold to accept the training data; this allows more false acceptances. When the cubic is replaced by a more general model of curves, these false predictions should improve. Figures 5-4 and 5-5 show more complex movements that are poorly represented by cubics. They also show a problem in inverse kinematics: the two curves are similarly shaped but offset. This is because small errors in location of a reflector on the dancer's body cause offsets in joint angle measurements. Step 8 is double-predicted on one dancer. This is due to missing data during the gap. Finally, the predictor for step 8 falsely accepted step 7 for one dancer. This is because the final parts of the steps sweep through the same range of angles, though at different velocities. Evaluating the labeling, there are 18 correct

True Step	Predictor output	Pred. errors	Timing errors start	end
plié	plié	0	.017	0
relevé	relevé	0	.075	.125
tendu	tendu	0	.117	.058
dégagé	dégagé	0	.117	.083
fondu	fondu, frappé	1		
frappé	fondu, frappé développé	2		
développé	frappé, développé g. batt. side	2		
g. batt. side	g. batt. side (double)	1		
g. batt. front	g. batt. front	0	.092	.067

Table 5.1: Scoring of test of two sequences of nine ballet steps, showing predictor errors and timing errors (Timing errors are not reported unless prediction is correct and unambiguous).

labelings, 5 false predictions, and 3 double predictions.

The learning algorithm was allowed to use only the seven variables shown in figure 5-2, which include Z-height scaled for the height of the dancer, one full set of Euler angles for the right hip and one extra hip angle from a different representation, and the right knee and ankle. Since the right leg was the working leg these are reasonable angles to consider, but why eliminate other variables? Several experiments were tried using 16 variables including six from each leg, Z-height, and torso orientation. This produced better results (confusion occurred only when the predictor for step 7 falsely labeled step 6); however, we believe the better results were due to a spurious correlation because of the small amounts of testing data, and that the effect is not robust. One of the learned predictors of tendu is also essentially a spurious correlation, as can be seen in Figure 5-6. The change in both variables is small; the predictor learned a patch in phase space rather than a curve.

A demonstration of invariance to speed and extension can also be seen in Figure 5-2. The two plots have different time bases: one is 16 sec., the other 22; an average of 50% difference in speed of movements. For Tendu the speed difference was 100% (a factor of 2). There is a difference of extension of 15% on relevé and 18% on dégagé. Despite these differences in speed and extension, the system showed excellent recognition on these steps.

Segmentation can be seen in the errors in the begin and end time estimates for steps 1, 2, 3, 4, and 9. The average error for those five steps is 67 msec, the maximum error is 150

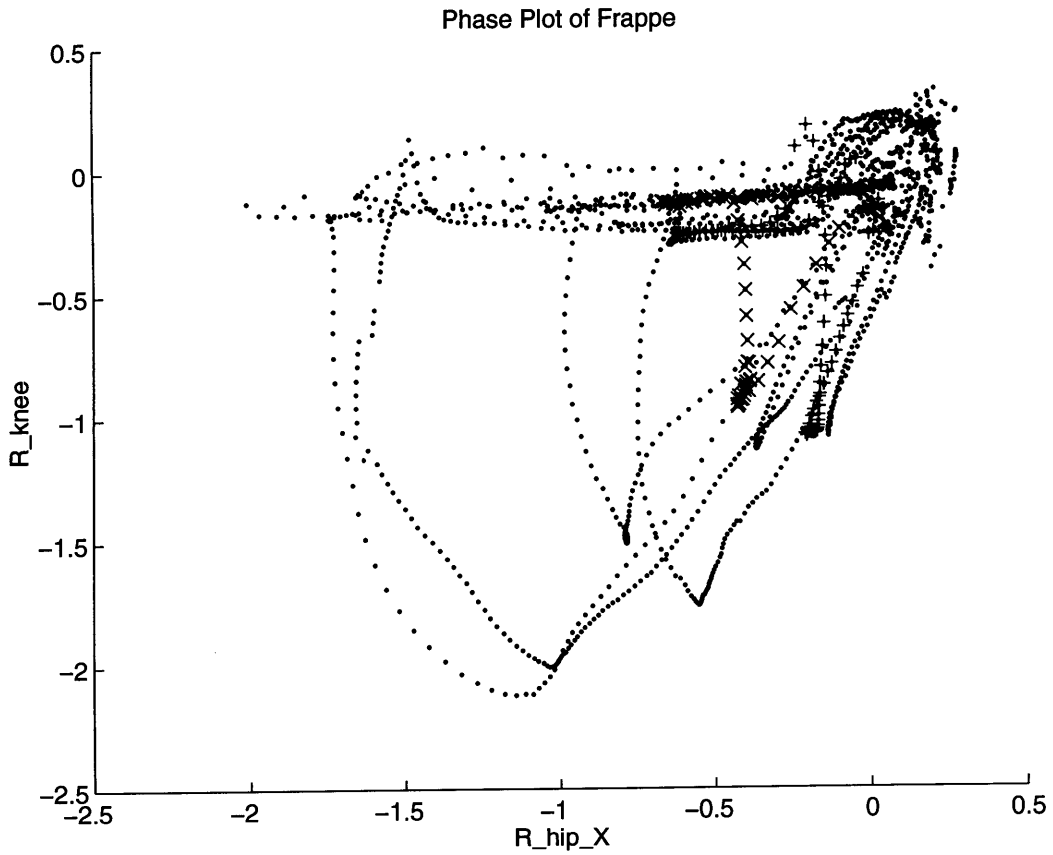

Figure 5-4: $\times$ and $+$ mark time steps during frappes for two dancers; "." marks time steps during other movements. This figure illustrates curves are more complex than cubics. The offsets are due to differences in angle measurement due to differences in reflector placement.

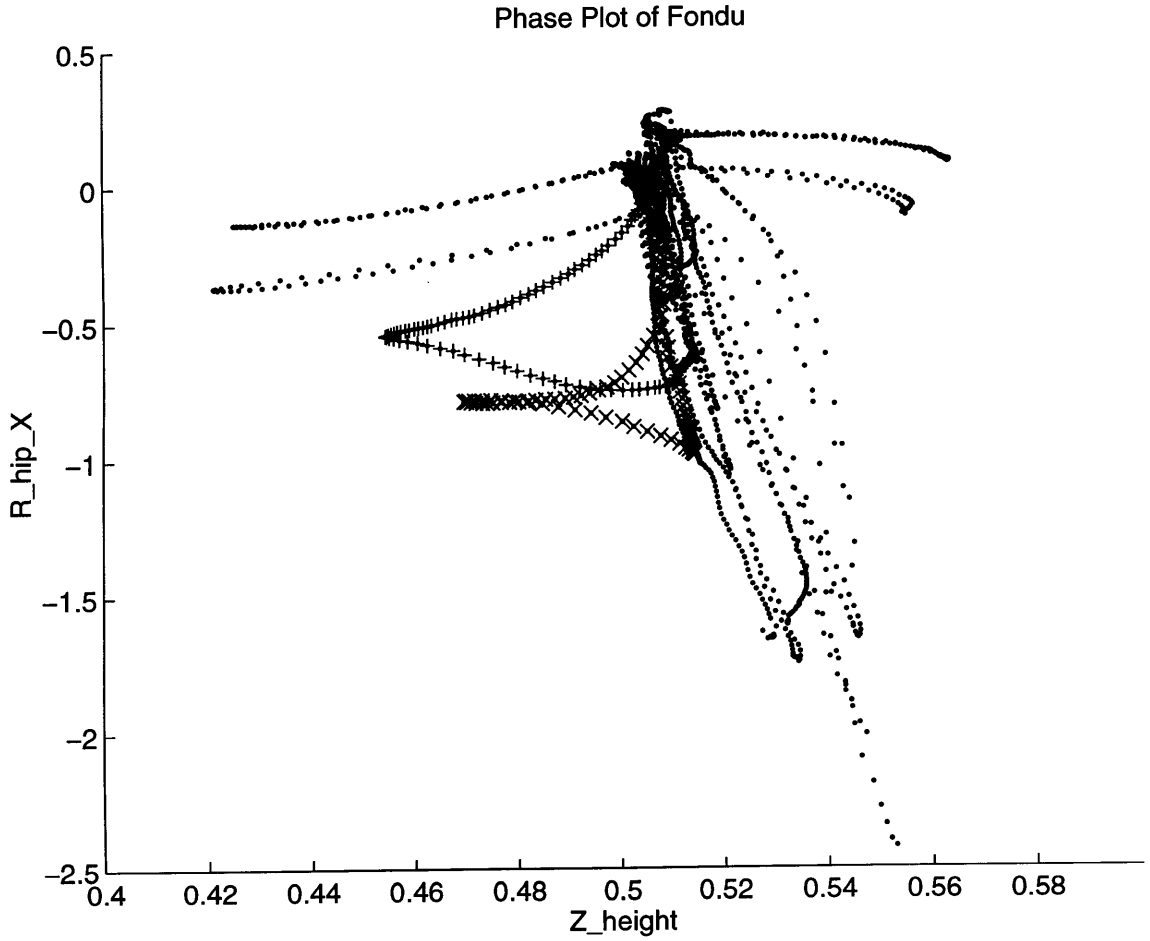

Figure 5-5: $\times$ and $+$ mark time steps during fondu for two dancers; "." marks time steps during other movements. This figure also illustrates more complex curves and offsets due to differences in reflector placement.

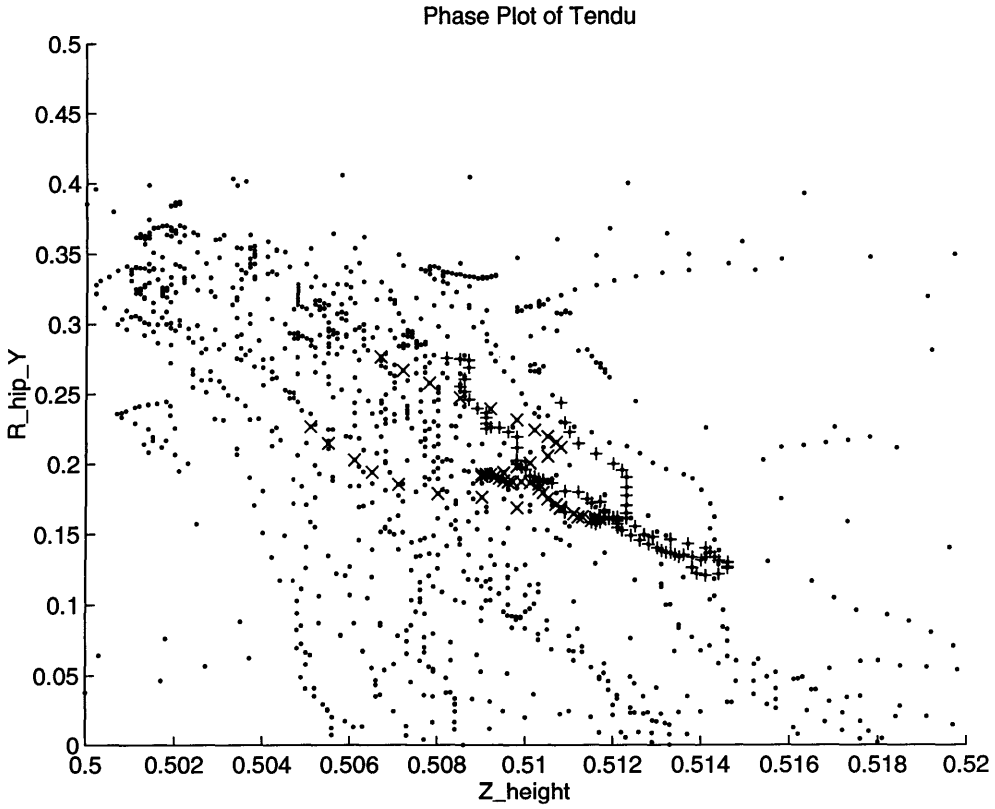

Figure 5-6: $\times$ and $+$ mark time steps during tendus for two dancers; "." marks time steps during other movements.

msec, and the minimum error is zero. From the start of data to the end of step 4 for the two sequences shown, there are 357 timesteps with 35 misclassified for one dancer, and 503 with 36 misclassified for the other. This is an error rate of 8% in classifying timesteps over the first four dance steps.

5.3 Summary of Results

The system meets its goals for ease of training: it trains automatically, and is robust to variations of 50% or more in the smoothing time constant and in the fitness function penalties. However, the system is less than halfway to its discriminability goals; discriminability should be improved when the model for atomic movements is improved from the present cubic curve. The system demonstrated abstraction over two fairly different dancers; more data from more dancers will be needed to better evaluate discriminability. The representation demonstrated invariance to speeds differing by up to 100%, and to extension differing by 18%. The system segmented more than half the steps with high accuracy. Noise tolerance was not directly tested; however it shows up in the system's ability to reject the noisy hip-Z data, and in the fact that the cubic model works as well as it does: since the cubic does not fit the data perfectly, the actual variations in the data show up as noise on top of the cubic model, and yet the system recognizes the actual data.

Clearly, the system needs to have the cubic model improved, and it needs more testing on more data. Despite such shortcomings, it demonstrated ease of training, abstraction, and segmentation on a difficult set of movements: tendu and dégagé are similar; fondu, frappé, and développé start out similarly, and tendu, dégagé, développé, and grand battement all end similarly. Adding turns and jumps to the test set should not add to the difficulty of predicting these nine steps because turns and jumps have characteristics not shared with these nine.

Chapter 6

Summary and Future Work

6.1 Summary

In this work we have developed a system for recognizing classical ballet steps from an input of XYZ tracking data. This problem is a part of the larger problem of understanding human body motion from images. We have assumed a solution will be found to the problems of finding and tracking people in images, leaving us free to work on issues of understanding. Here our working definition of "understanding" is identification: translating from a signal representation to a symbol representation where the symbols are at the level of detail natural to human thought and communication.

Our examination of representation issues led us to a novel "phase space" representation based on techniques from classical physics. Our review of the literature shows other workers using low dimensional phase spaces of position and velocity, and no other workers projecting out all the velocity axes to achieve speed independence. In our phase space, the axes are joint angles, points represent configurations of the body, and movements are space curves through a continuous sequence of body positions. Among the claimed advantages to the representation were invariance to changes in speed and extension. In testing, the representation demonstrated invariance to differences of up to 100% in speed and 18% in extension.

We developed a learning method that examines relationships between pairs of joint angles, searching for those relationships which best predict category membership of ground truth data, and constructing simple detectors from the relations. The method then compounds the best detectors, which improves detector performance. The learning method

works automatically on movement sequences annotated with ground truth ballet step identifications. There are two parameters to the learning algorithm which must be set manually: the penalty weight ratio for the fitness function which evaluates predictors, and the smoothing time constant. The learning system shows robustness to variations in both of these parameters, producing similar results for changes of 50% in each of the parameters.

The ballet step recognition system works extremely well on five out of the nine test ballet steps, correctly identifying them and estimating the start and stop times with an average error of only 67msec. On the other steps it sometimes produced two identifications of the step; one correct and one incorrect. These errors point the way for future work.

6.2 Future Work

The testing performed so far has demonstrated the plausibility of phase space recognition, but not yet its viability. To show viability, we must improve the system so it can do higher accuracy prediction then demonstrated so far, and on a larger number of ballet steps. It is also important to test the representation on different kinds of data. Below we propose some improvements, other applications, and extensions.

At present, pair relations are represented by simple cubic curves, and this is adequate to represent simple movements of plié, relevé tendu, dégagé, and grand battement. However fondu, frappé, and développé are actually composed of two simple movements and thus are not well represented by simple cubics. One approach to this problem is to rigorously define an "atomic" movement as one representable by simple low order curves, and enforce that definition when annotating ground truth data. However, doing so deviates from the principle of identifying movements at a level of detail natural to humans. A preferable approach is for the system to detect compound movements of a small number of segments, and separate them into their simple components automatically.

The system could be improved by considering velocities. As long as the system processes only positional variables such as joint angles, it is evaluating movement based on kinematics. The smoothing filter ensures a degree of continuity of positions which goes beyond pure kinematics, however processing velocities would allow it to evaluate dynamics. Because many movements occur at varying speeds, it may make sense to consider velocity inequalities, e.g. the system might learn that velocity is negative in one part of the plié and

and positive in another part.

The representation developed in this work may have applications outside of human body movement. We intend to apply it to facial expression recognition, building on the work of Essa [16]. Essa tracks facial movements from images, and extracts estimates of muscle activation. We intend to use these muscle estimates as axes of a phase space, and see if, when different subjects make the same expression, their muscle activations map into the same region of phase space.

Another possible extension of this system is to try and do recognition on 2D moving light displays of dancers instead of 3D tracking data. To do this we would assume weak perspective and the camera pointing approximately horizontal. Most researchers assume this, and they also assume fronto-parallel motion. If that assumption is relaxed, we are left with one degree of freedom, which can be represented as the facing direction of the dancer. It may be possible to search this one degree of freedom and still recognize ballet steps.

The contribution of this work is the use of the phase space representation for both representation and recognition of human body movement. Phase space has been widely used in physics for decades, and in the computer vision community it was used by Shavit [41] for recognition. We extend that work by finding subspaces of phase space that contain invariants of the motion (our subspaces are formed by a subset of the axes of the original space). We also show that it is possible to segment movements (detect start and end times) using phase space.

Appendix A

The Tracking System and Data Gathering

Our tracking data came to us by the courtesy of Adaptive Optics Associates of Cambridge MA, who kindly loaned us their MultiTrax cameras, electronics, and software, as well as their expertise in using the equipment.

The Multitrax system does video-based tracking. Each camera lens is surrounded by infrared LED's, and the camera lenses have filters to make them sensitive to IR only. The dancer is fitted with reflective spheres or hemispheres about 4cm diameter on toes, ankles, knees, hips, shoulders, elbows, and wrists. The reflectors use tiny transparent beads which reflect light strongly back to its source, similar to highway signs. The video signal from the cameras goes directly to special processors which extract target centroids in real time. The video processors are all connected by serial links to a Macintosh computer running software to store the data from each camera. We used two, four, and six cameras in various runs, and up to twenty reflectors on the dancer (in some cases head and hands were tracked for other users of the data).

Reflector placement can cause two kinds of problems in data gathering. One is that a reflector can be occluded from one or more cameras by some part of the dancer's body. This problem is worst with the ankles. The other is that one reflector can come too near another in some camera's view. Then the two targets merge into one and a single "average" centroid is recorded, which produces erroneous 3D data. We tried to be consistent in placing reflectors on the dancers. We wanted unhindered dance, we wanted the reflectors visible as

much of the time as possible, and we wanted the reflectors as near the joints as possible; not all of these conditions could be met.

We had the most problems near the ankle reflectors, because fifth position presses the inside of one ankle very near the outside of the other, and many motions involve brushing the feet past the ankles. We settled for hemispheres placed on the inside just forward of the projecting ankle bones. This had the disadvantage of poor visibility in fifth position, but the advantages were: good separation in first position and relative freedom to brush the backs of the ankles. Toe reflectors were placed on top of the foot opposite the ball of the foot. Knee reflectors were placed on the inside between the kneecap and the top of the tibia. Hip reflectors were placed just above the crease between torso and leg, and centered on the width of the leg. Shoulder reflectors were placed on top of shoulders where the clavicle meets the scapula. Elbow reflectors were placed on the outside of the elbows on the end of the radius. Wrist reflectors were placed on the back of the wrists just before the wrist joint. When hand reflectors were used, they were splinted to the middle fingers so they stayed at the fingertips. When head reflectors were used, hemispheres were placed above the outside corner of each eye, and a spherical reflector on top of the head towards the back. We used two-sided medical electrode tape to adhere the reflectors to the dancers' clothes, and we placed their head reflectors on scarves. Only the wrist reflectors were attached to skin.

The cameras are set up to view the sensing area. They are calibrated by placing in the sensing area a calibration frame with nine reflectors at known locations, and orienting it so each camera can see all reflectors simultaneously. This calibration frame determines the coordinate system of the tracking data, and the orientation of each camera in that coordinate system.

When data files are saved, they are saved with calibration data, and with 2D target centroids from all video processors. An offline, semi-automated process called "sorting" establishes correspondences between camera views and tracks points over time.

We gathered data in several different sessions using different numbers of cameras, and alternating between two dancers. each run was from 10 to 25 seconds and contained five to ten dance steps. Sometimes the steps were done with deliberate pauses between them; other times they were chained together in a flowing sequence. We collected more data than we used in this thesis.

The sorting process can automatically track points, but it requires human operator

intervention to establish the initial correspondences and sometimes to re-establish correspondence after a reflector disappeared and reappeared. The operator must also be vigilant in detecting and cancelling merged markers. Otherwise a bad centroid could be used, resulting in bad 3D data. A later version of the tracking software alleviated many of the sorting problems.

The sorted tracking data consists of the XYZ coordinates of each marker, or missing data codes when a marker is obscured. Our system begins by extracting joint angles from the tracking data according to a human body model with 1DOF joints at ankles, knees, and elbows, and 3DOF joints at hips and shoulders, plus six more coordinates for torso location and attitude.

For each time step where there is missing XYZ data for a marker, joint angles associated with that marker cannot be computed, so missing data codes are propagated to the uncomputable angles. For example if there is missing data for a hip marker, then both hip and knee angles are flagged as missing. The ankle markers are the most frequently obscured; when only an ankle marker is missing, both ankle and knee are marked missing and at most two of the three hip angles can be computed depending on the representation. In those cases the system computes data for the angles determined only by the knee position, and marks the rest as missing data.

Bibliography

[1] Harold Abelson and Gerald Jay Sussman. The dynamicist's workbench: I automatic preparation of numerical experiments. Technical Report 955, MIT Artificial Intelligence Lab, 545 Technology Sq., Cambridge MA, 1987.

[2] Koichiro Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, 17(1):73–83, 1984.

[3] Mark Allman and Charles R. Dyer. Towards human action recognition using dynamic perceptual organization. In *Looking at People Workshop, IJCAI-93*, Chambery, Fr, 1993.

[4] Minoru Asada and Saburo Tsuji. Representation of three dimensional motion in dynamic scenes. *Computer Vision, Graphics, and Image Processing*, 21(1):118–144, Jan 1983.

[5] N. I. Badler and S. W. Smoliar. Digital representation of human movement. *ACM Computing Surveys*, 11:19–38, March 1979.

[6] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23(2):145–152, 1978.

[7] Joan Benesh. *Benesh Dance Notation*, volume 1-4. College of Choreology, London, 1967.

[8] T. O. Binford. Visual perception by computer. In *IEEE Conference on Systems and Control*, Miami, Fla, 1971.

[9] Aaron Bobick. Natural object categorization. Technical Report 1001, MIT Artificial Intelligence Lab, 545 Technology Sq., Cambridge MA, 1987.

[10] Aaron Bobick. Representational frames in dynamic scene annotation. In *Proceedings of IEEE Workshop on Visual Behaviors*, Seattle, WA, 1994.

[11] Richard A. Bolt and Edward Herranz. Two-handed gesture in multi-modal natural dialog. In *Proceedings of UIST '92, Fifth Annual Symposium on User Interface Software and Technology*, Monterey, CA, 1992.

[12] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*, chapter 1. Wadsworth & Brooks, Pacific Grove, CA, 1984.

[13] Winand H. Dittrich. Action categories and the perception of biological motion. *Perception.*, 22(1):15, 1993.

[14] Richard Duda and Peter Hart. *Pattern Classification and Scene Analysis*, chapter 3. John Wiley, New York, 1973.

[15] Noa Eshkol. *The quest for T'ai Chi Chuan.* Eshkol-Wachman movement notation. Research Centre for Movement Notation at the Faculty for Visual and Performing Arts, Tel Aviv University, Tel Aviv, Israel, 2nd and expanded edition, 1988.

[16] Irfan Essa, Trevor Darrell, and Alex Pentland. Tracking facial motion. Technical Report 272, MIT Media Lab Vision and Modeling Group, 20 Ames St, Cambridge MA, 1994. To appear in IEEE Workshop on Nonrigid and articulated Motion, Austin TX, Nov 94.

[17] Nigel Goddard. *The Perception of Articulated Motion: Recognizing moving light displays.* PhD thesis, University of Rochester, June 1992.

[18] Kristine Gould and Mubarak Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characteristics. In *Proc. 1989 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 79–85, 1989.

[19] Gail Grant. *Technical Manual and Dictionary of Classical Ballet.* Dover, New York, 1967.

[20] Judith A. Gray, editor. *Dance Technology: Current Applications and Future Trends.* The American Alliance for Health, Physical Education, Recreation, and Dance, Reston, VA, 1989.

[21] Ann Hutchinson Guest. *Labanotation The System of Analyzing and Recording Movement.* Routledge, Chapman and Hall, New York, revised third edition, 1977.

[22] H. Haken, J. A. S. Kelso, A. Fuchs, and A. S. Pandya. Dynamic pattern recognition of coordinated biological motion. *Neural Networks*, 3:395–401, 1990.

[23] Jessica K. Hodgins. Biped gait transitions. *IEEE Transactions on Robotics and Automation*, 7(3), 1991.

[24] D. D. Hoffman and B. E. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42:195–204, 1982.

[25] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, Feb 1983.

[26] Peter C. Hughes. *Spacecraft Attitude Dynamics*, chapter 2. John Wiley, New York, 1986.

[27] E. A. Jackson. *Perspectives of nonlinear dynamics*, chapter 2. Cambridge University Press, 1989.

[28] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[29] Gunnar Johansson. Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.*, 38:379–393, 1976.

[30] G. A. Korn and T. M. Korn. *Mathematical Handbook for Scientists and Engineers*, chapter 14.10. McGraw–Hill, New York, 1968.

[31] D. Marr and H. K. Nishihara. Representation and reconition of the spatial organization of three dimensional shapes. *Proceedings of Royal Society of London B*, 200:269–294, 1978.

[32] P. Morasso and V. Tagliasco, editors. *Human Movement Understanding.* Elsevier Science Publishers B. V. (North-Holland), Amsterdam, 1986. see Chapter 3 on Movement Notation by A. Camurri et al.

[33] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. Technical Report 290, MIT Media Lab Vision and Modeling Group, 20 Ames St, Cambridge MA, 1994. To appear in IEEE Workshop on Nonrigid and articulated Motion, Austin TX, Nov 94.

[34] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-2:522–536, 1980.

[35] R. Polana and R. Nelson. Detecting activities. In *Proc. 1993 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 2–7. IEEE Press, 1993.

[36] Krishnan Rangarajan, William Allen, and Mubarak Shah. Matching motion trajectories using scale space. *Pattern Recognition*, 26(4):595–610, 1993.

[37] R. Rashid. Towards a system for the interpretation of moving light displays. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2:574–581, 1980.

[38] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, Jan 1994.

[39] Elisha Sacks. Automatic qualitative analysis of dynamic systems using piecewise linear approximations. *Artificial Intelligence*, 41:313–364, 1989-90.

[40] Stan Sclaroff. Physically-based combinations of views: Representing rigid and nonrigid motion. Technical Report 273, MIT Media Lab Vision and Modeling Group, 20 Ames St, Cambridge MA, 1994. To appear in IEEE Workshop on Nonrigid and articulated Motion, Austin TX, Nov 94.

[41] Eyal Shavit. *Phase Portraits for Analyzing Visual Dynamics.* PhD thesis, University of Toronto, Jan 1994.

[42] Eyal Shavit and Allan Jepson. Motion understanding using phase portraits. In *Looking at People Workshop, IJCAI-93*, Chambery, Fr, 1993.

[43] Thad Starner, John Makhoul, Richard Schwartz, and George Chou. On-line cursive handwriting recognition using speech recognition methods. In *ICASSP 94*, 1994.

[44] Valerie Sutton. *Dance writing shorthand for classical ballet.* Sutton Movement Writing Press, 1981. Also see: The Movement Shorthand Society, Irvine, CA.

[45] Shimon Ullman. *The Interpretation of Visual Motion*, chapter 4. MIT Press, Cambridge, MA, 1979.

[46] Agrippina Vaganova. *Basic Principles of Classical Ballet, Russian Ballet Technique.* Dover, New York, 1969. Translated from the Russian by Anatole Chujoy.

[47] A. Verri, F. Girosi, and V. Torre. Mathematical properties of the two-dimensional motion field: from singular points to motion parameters. *Journal of the Optical Society of America*, 6(5):698–712, May 1989.

[48] Video Dictionary of Classical Ballet . 4 hours on 2 Videotapes, 1990. Distributed by: Kultur International Films LTD, West Long Branch, NJ.

[49] Rudolf von Laban. *Laban's dance notation charts : basic symbols.* Associated Music Publishers, New York, 1928. Repubished in 1975 by Macdonald & Evans, London.

[50] Gretchen Ward Warren. *Classical Ballet Ballet Technique.* Univ. of S. Fla Press, Tampa, 1989.

[51] J.A. Webb and J.K. Agarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19:107–130,, 1982.

[52] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequetial images using hidden markov model. In *Proc. 1992 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 379–385. IEEE Press, 1992.

[53] Jianmin Zhao. *Moving Posture Reconstruction from Perspective Projections of Jointed Figure Motion.* PhD thesis, University of Pennsylvania, Aug 1993.