M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 309
Abbreviated version appears in Fifth International Conference on Computer Vision, Pp. 624-630, Cambridge MA, 1995

Recognition of Human Body Motion Using Phase Space Constraints

Lee Campbell **Aaron Bobick**
Room E15-383, The Media Laboratory
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139 *
elwin@media.mit.edu bobick@media.mit.edu

Abstract

A new method for representing and recognizing human body movements is presented. Assuming the availability of Cartesian tracking data, 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 axes of 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 constraints which are in effect during the movement to be learned, and not in effect during other movements. We then use the learned representation for recognizing movements in data.

Prior approaches by other researchers used a small number of classification categories, which demanded less attention to representation. We train and test the system on nine fundamental movements from classical ballet performed by two dancers. The system learns and accurately recognizes the nine movements in an unsegmented stream of motion.

1 Introduction

Until recently, computer vision analysis of image sequences focused predominantly on issues of geometry, either the three-dimensional geometry of a scene or the geometric motion of a moving camera (see [18] for review). Lately, however, there has been an effort at getting computers to understand action or behavior [7]. In this paper we present a technique for recognizing human motion in domains in which there are clearly defined semantic categories of movement.

There are numerous domains — athletics, dance, surveillance — in which the action or behavior of objects is much more semantically rich than the static configuration of the scene components. To bound the scope of the problem, to establish performance criteria, and to be certain that we are extracting semantic categories that are meaningful to people we require that the domain already have a well defined vocabulary for describing action. Since people use that vocabulary for describing the action we know that such a description captures the important relevant semantics. If it did not, the vocabulary would not provide a reasonable basis for communication. Finally, a well defined vocabulary reduces the difficulty in establishing ground truth; people should find it easy provide a baseline against which a system could be tested.

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.

The description of human motion is one component of the general problem of *video annotation.* Entertainment companies, newscasters, and sports teams are acquiring ever larger video databases. If these can be (semi-)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 using a vocabulary in which people naturally describe the domain. One attractive aspect of this scenario is that the speed of annotation is relatively unimportant.

In this paper we present a method for recognizing classical ballet steps from three-dimensional point data. We are using 3D data as input, as opposed to video, because we wish to focus on the recognition task, not the geometry recovery. There has been much previous work on recovering 3D models from 2D motion; our question is given 3D data, how do you see a plié? The approach we develop is based on the idea that different categorical movements each embody a different set of constraints on the motion of the body parts; these constraints are mostly easily observed in a phase-space that relates the independent variables of the body motion. By learning which sets of constraints are highly diagnostic of particular motions we can build constraint set detectors to recognize the movements.

2 Symbolic understanding of human motion

The problem of understanding human body motion from images is one that leads into such diverse areas as dynamics, athletics, and cognitive science. Since our primary motivation is video annotation, we focus on a symbolic description of action that translates the continuous domain of human motion into a discrete sequence of symbols.

2.1 Constrained human motion: ballet as example

With few exceptions (e.g. finger manipulation) human body motion has many constraints. Some of the constraints come from the laws of physics: arms do not separate at the shoulder and people need to maintain balance to stay off the ground. Other constraints, which come from the rules of an athletic or artistic form, we will call cultural constraints. The fundamental idea of the work presented here is that by recognizing the presence of these constraints it is possible to recognize the motions.

Gaits are an example of highly constrained motion. Physical constraints induce a rhythmic and repetitive pattern of motion [23].

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

From a computational vision perspective these constraints are so strong that several researchers treated walking as a periodic 1DOF motion for purposes finding the limb orientations [17, 28], or the identity [24] of a walking person.

The domain of ballet

Ballet is a domain rich with both 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 and each of a collection of experts typically will generate identical descriptions. In addition to the named steps there are classifications or categories of steps and written notation languages for recording ballet and other forms of dance [15, 22]. The good agreement among experts indicates that the moves are quite distinct and form a reasonable set for a computer to try to discriminate.

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 [14, 32].

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.

The physical constraints of ballet are the most obvious ones. As in all well controlled motions, maintenance of support and balance is essential. 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 from outside 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. Analogously, when two dancers execute the same ballet step, the physical constraints require a high degree of similarity.

Ballet is also subject to numerous cultural constraints which increase the similarity of step from one execution to the next, from one dancer to the next. "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, and are well defined for each move. The execution of the steps is also subject to the constraint of "grace:" it is better to do a movement with less extension gracefully than with large extension that shows straining. The set of physical and cultural constraints result in highly repeatable movements.

Of course ballet steps are not executed exactly the same way each time; if they were, standard recognition techniques based on a complete spatio-temporal geometrical description could be exploited. When steps occur in sequences, the beginnings and ends are modified to transition to the neighboring step. This is analogous to co-articulation in speech. Another source of variation is called "extension:" 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.

2.2 Defining the problem

As mentioned in the introduction, many researchers are working on recovering the three-dimensional geometry of moving people (e.g. see [9, 28]). For the work here we assume that three-dimensional position of some body points are available and the goal is to recognize the ballet steps from that data. Our system uses a commercially available system [3] to provide video-rate 3D data of approximately 20 markers attached to a human body.

Our specific task is to learn and identify a set of nine atomic ballet movements from the XYZ tracking data. The nine movements, all performed starting from a flat-footed position with legs turned out and with the right leg working and the left leg supporting, are shown in Figure 1 and described below:

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);
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;

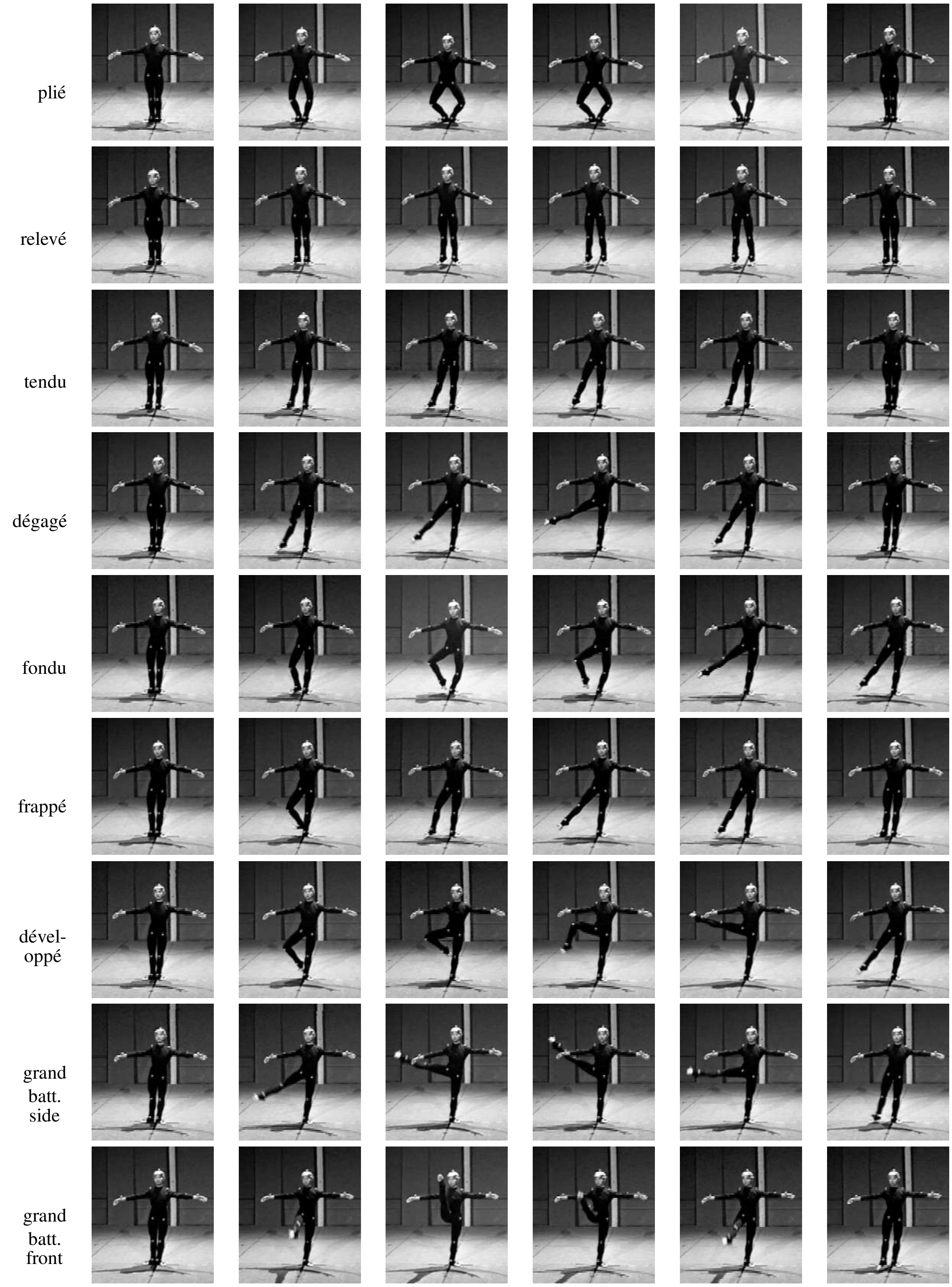

Figure 1: Nine "atomic" ballet moves.

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

We will use two measures of success. The first is a standard pattern recognition evaluation function where, given ground truth, we measure the percentage of time instances in which the system correctly identifies dance step in progress. A more relevant performance measure is the ability of the system to generate the correct annotation, announcing the same sequence of symbols (dance steps) that a human observer would generate.

3 Related work

In describing related work we focus on (1) the psychophysics and structure-from-motion work that demonstrates the plausibility of the solving the specific problem of recognizing motion from the position of 3D points; and (2) previous attempts at recognizing specific motions (for a more complete discussion of related work see [8]).

In 1973 Gunnar Johansson [19] published results from a series of psychophysical experiments which proved that people can recognize movements of humans from Moving Light Displays (MLDs) — 2D images of a small number of moving spots attached to the joints. Later results (e.g. [20, 5, 10]) demonstrated the ability of humans to recognize movements such as hammering, box lifting, ball bouncing, and stirring, and two-person activities such as dancing, greeting and boxing.

Early computational vision results (e.g. [31]) proved basic structure-from-motion theorems demonstrating it is possible to recover the three-dimensional geometry behind MLD's. Hoffman and Flinchbaugh [16] and Webb and Aggarwal [33] exploited the fact that biological motion mostly involves 1DOF rotation joints to compute the three dimensional motion of restricted human movement. More recent work [35, 12] have focused on exploiting a model in the interpretation of MLD's.

The psychophysics results combined with the 3D-geometry methods raises the issue of whether people perform the task of motion recognition from MLD's using two-dimensional or three-dimensional information. Subjects could either recognize the motion directly from the image positions, or first compute 3D information, and then recognize the movement. However, because the 2D input is trivially computable from the 3D information, the results certainly imply that the task should be easy to perform given the 3D position of sufficient points.

Understanding body motion

O'Rourke and Badler [25] tracked features in synthetic imagery, using constraint propagation to implement a feedback loop between high-level and low-level processes of image motion analysis. More recently [17, 1, 28] research has been focused on on determining correspondence between frames of an image sequence and the phase of a known body motion. The use of a model with known motion (and typically a known background) greatly simplifies the problem as certain key features, such as the first few principal components of a shape contour, can be used as feature vector and can be matched against known vectors.

Additional efforts have devised methods of representing general (model-free) motion. Rangarajan, Allan and Shah [27] developed a system to match trajectories using scale space images of curves of speed vs time and direction vs time. Polana and Nelson [26] present a textural method for detecting periodicity in an XYT solid and, assuming a stationary camera, matching to other periodic motions and rejecting non-periodic motions. Allman and Dyer [2] find optic flow in an XYT solid, trace curves as a function of time, and cluster the curves. Gould and Shah [13] find features on curves from two representations based on trajectories: one curve is X and Y velocity vs time; the other curve is distance vs local curvature. These feature sets can then be compared to previously represented motion.

Perhaps the work most similar to that presented here is that of Shavit [29, 30]. This technique represents "qualitative visual dynamics" using a phase space, the classical technique for analyzing the dynamics of a system. Phase space is a Euclidean space with axes for all the independent variables of a system and their time derivatives. 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 autonomous systems — systems in which parameters such as spring constants do not change with time and to which energy is not being added — only one phase path passes through each point in the space. In such systems, knowledge of position and velocity of each variable at one time instant completely determines the behavior of the system. Simple dynamic systems (e.g. mass on a spring, or viscous damping) have easily characterized phase portraits in phase space.

In Shavit's work, a binary region in the image is described as a blob of a deformable elastic material. Over time the blob deforms, and the deformations are represented by a time varying deformation parameter vector. The image motion is then described as a trajectory through the phase space that is a function of these parameters. Using training examples of images of known motion, Shavit constructed phase diagrams, and segmented them into sections with simple dynamics. Finally, imagery such as black silhouettes of walking, running, and jumping cartoon characters, moving light displays of actors performing the same gaits, and images derived from films of running animals were characterized by in terms of combinations of these simple dynamic sections. With these characterizations Shavit was able to identify characteristics about the gaits which are suitable for identification of the gaits. No attempt has yet been made to identify a larger set of movements or motion classes.

Before concluding our discussion of phase-space methods we note that in non-autonomous systems specifying the position and velocity of the independent variables at one time instant is *not* adequate to describe the motion through phase space. Time varying parameters require that the values of the parameters be specified at many time instances; in the limit one would provide complete parameter trajectories as a function of time. In such cases, specifying velocity information is less critical; approximate values of the velocity can be inferred by finite differencing the time sampled parameter trajectories.

Other methods: Hidden Markov Models (HMM's) were used by Yamato et al. [34] to recognize images of tennis strokes seen from a particular viewpoint. A typical pattern classification technique such as Maximum Likelihood [11] would require a sequence of movements to be segmented into separate dance steps before the steps could be classified. All the difficulty in this approach becomes concentrated in the segmentation phase.

4 Representation and Recognition in Phase Space

Central to this work is the problem of how to represent movement. Two kinds of considerations bear on the choice of representation. One is intrinsic, e.g. is it canonical and can it be stably computed from raw data. The other, extrinsic, involves use of the representation — in this case, how well does it support recognition. In this section we will briefly motivate our choice of representation, explicate its details, and consider its intrinsic properties. Next, we develop a recognition method based upon the representation that requires learning the best parameters to describe each motion. In section 5 we present the results of our recognition experiments.

4.1 Representation Details

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 a full phase space of all the torso position and attitude parameters and joint angles, and all their derivatives. Plotting only the leg variables while holding the arm variables constant would show a space curve, and each repetition of the movement would traverse approximately 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 — the set of constraints in effect during its execution — is captured in the leg variables, while the variation is in the arm parameters: the constraint of the motion can be found in a subspace of the full phase space.

Figures 2 and 3 illustrate this idea. Each figure shows a 2D-projection of phase space in which a collection of data points are plotted. The data points corresponding to two dancers performing a plié are marked by $\times$'s and $+$'s; all other data points are marked by a "$\cdot$" Note that the points corresponding to the plié and the relevé are nicely segregated from the other data points. This intuition that the constraints in effect during a motion are visible in a subspace of phase space motivates our choice of representation for movement.

Define "body phase space" to be a space with axes for each independent parameter of the body model and their first derivatives but no independent time axis. Let a *2D-projection space* be a two-dimensional subspace of body phase space spanned by any two axes of the original phase space. Figures 2 and 3 each show two examples of 2D-projection spaces. We can specify 2D-projection space curves by defining curves in the plane of the 2D-projection space. And, finally, we represent motion by a collection of these 2D-projection space curves. If considered in the full phase space, the collection of these 2D-projection curves define an intersection of manifolds, each defined by one curve.

Continuing with the plié example, suppose we draw a curve through the $\times$'s and $+$'s in each of the two 2D-projection spaces of figure 2. If we loosely say that "currently performing a plié" is equivalent to the state variables being near both of the hypothesized curves, then, in the full phase space, we have defined the intersection of the R_knee/Z_scaled manifold with that of the R_ankle/Z_scaled to define a plié. We have also coarsely outlined a strategy for recognizing a motion: 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 each of the 2d-projection space curves used to define the motion. We postpone any further discussion of recognition until we after consider the intrinsic properties of the representation.

Before continuing we mention some of the specifics. For recognizing ballet steps we use cubic polynomials to form the 2D-projection curves. A low order curve implies that a single move cannot require too complicated a motion, agreeing with our intuition that a single action is "simple." Of course, any continuous curve parameterization is possible such as piecewise linear segments or b-splines. Also, for the examples shown, the body parameters considered are joint angles for 1DOF joints (e.g. knee), Euler angles for 3DOF joints (e.g. hip), torso orientation, and body-height-normalized torso height above the floor plane.

4.2 Representation considerations

Aspects of the general representation problem are discussed in Marr and Nishihara [21] and Badler and Smoliar [4]. Marr and Nishihara consider representations of three dimensional shapes for object recognition and present three criteria: accessibility, scope and uniqueness, and stability and sensitivity. These criteria, however, are as relevant to recognizing motions and we evaluate our representation in light of these considerations.

The representation is clearly *accessible*; it is easy to compute joint angles from the raw Cartesian tracking data. In our application, *scope and uniqueness* refer to the class of movements for which the representation is designed and whether movements have canonical descriptions in the representation. Since a point in body phase space completely determines the state of all body parts, two coincident phase paths represent the same motion. Furthermore, the phase variables are uniquely determined by the body geometry so that two identical motions map two identical phase paths.

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?

By selecting only a subspace of body phase space we can ignore dimensions that are not constrained by a movement. As mentioned, the arms can be excluded when assessing whether a move is a plié. Equally important, if temporal variation — variation in the speed at which a move is executed — is to be abstracted, then all the derivative axes of the body phase space should be excluded. In principle, the learning algorithm in Section 4.4 can learn to ignore all the derivative dimensions; for the results presented in Section 5 we explicitly removed from consideration the velocity axes of the phase space.

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 redundant sets of Euler angles to represent 3DOF joints, and to do represent a motion using the angle set which is not near singularity.[1] The recognition system described in the next section *automatically* selects the stable angle set to represent and recognize a given motion.

Quaternions were also considered for representing 3DOF joints. However, an 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 more robust in the presence of missing data.

[1] 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.

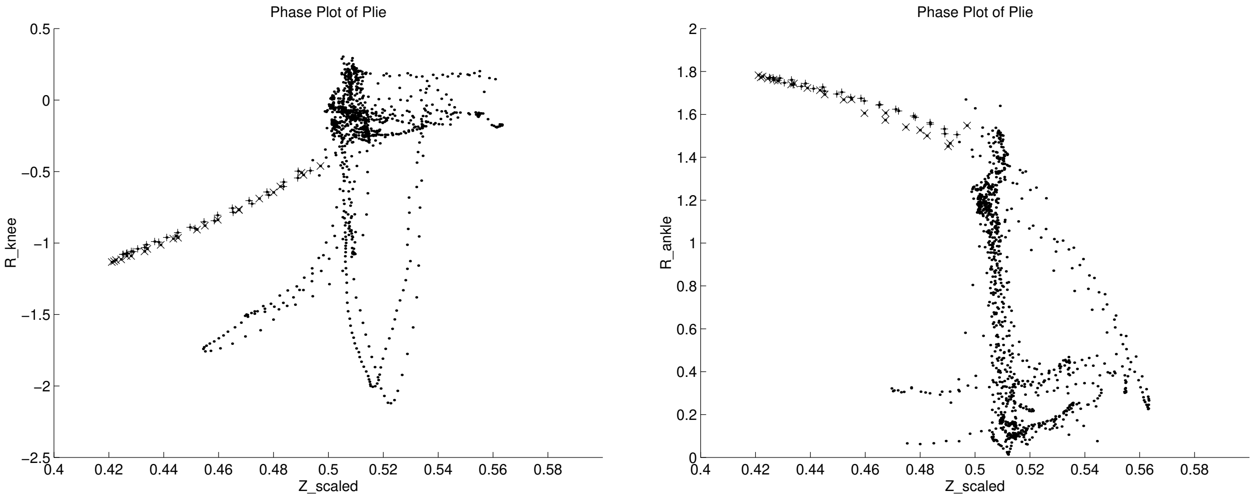

Figure 2: $\times$ and $+$ marks time steps during pliés for two dancers; "." marks time steps during other movements. Angles are in radians. These two views of the phase space show how the plié lies in a region separated from the other movements (data subsampled for clarity).

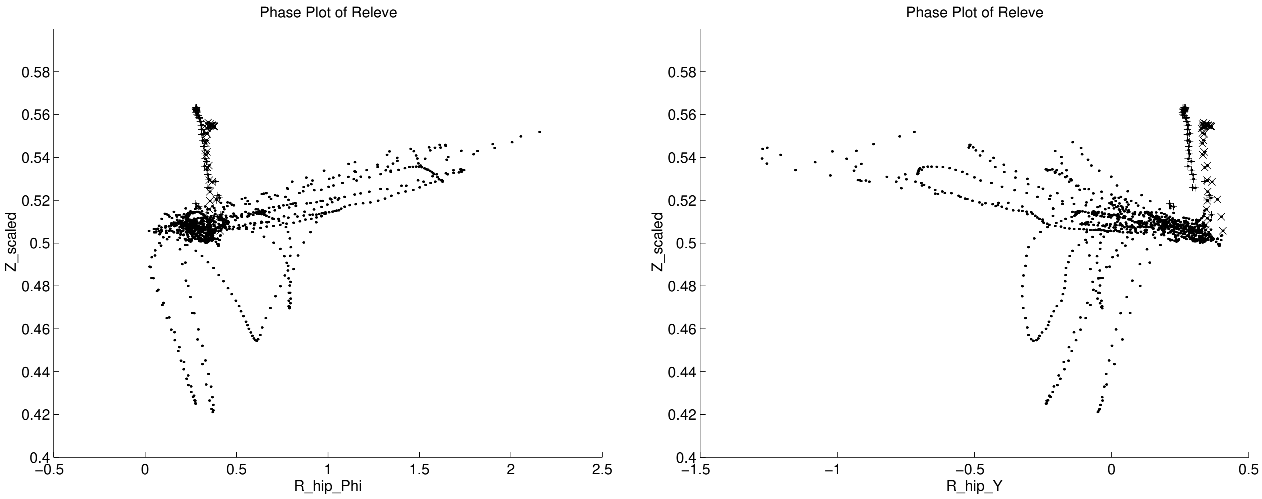

Figure 3: $\times$ and $+$ mark time steps during relevés for two dancers; "." marks time steps during other movements. Two views of relevé.

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.

4.3 Recognition in phase space

What does it mean to recognize a ballet step? From the perspective of annotation, recognition requires announcing the correct step somewhere during its performance, and not announcing any other steps. However, we begin the development of our recognition system by considering the identification of the currently executed move at each time step.

The basic idea is as follows. At each time instance the state of the system can be represented as a point in phase space. We will accept a point as being part of a recognized movement if it is within a 2D-projection-specific threshold distance of each of the 2D-projection space curves used to define the move. [2]

Let us define the following terms:

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. This is a curve lying in a 2D-projection space.

Threshold: a distance h above and below a pair predictor

[2] 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.

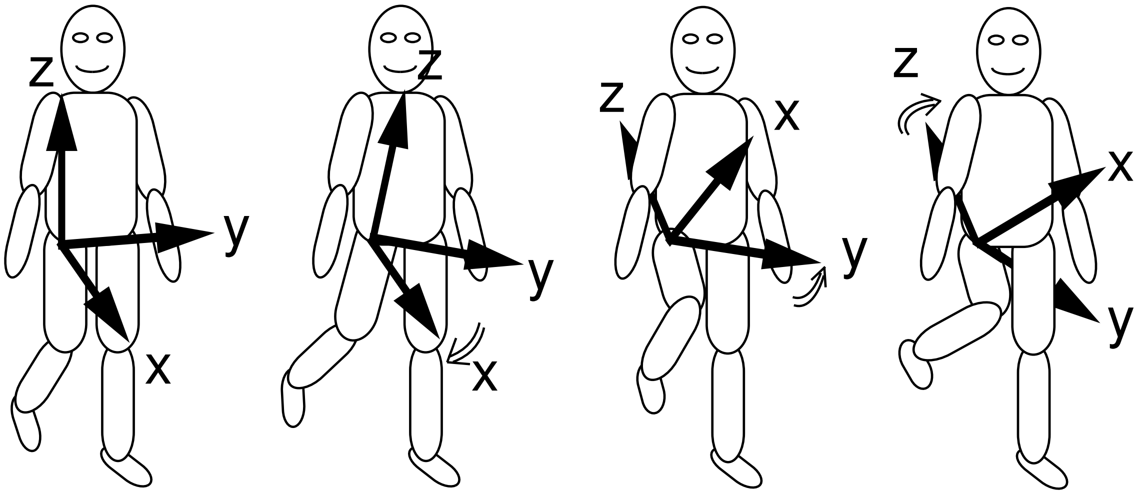

Figure 4: Three successive rotations illustrate Euler angles representing 3 rotational degrees of freedom of hip joint. Coordinate axes are rigidly attached to femur, i.e. rotations are measured in frame of reference of leg. If the leg is straight and toe pointed such that all 4 reflectors lie on a line, a singularity occurs and the Z angle cannot be recovered from data. Near this singularity the Z measurements suffer from excessive noise.

$|\psi_j - f_{ij}(\psi_i)| \leq h$ 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 pair relation and a threshold.

In the current implementation, a pair relation for a given movement in a given 2D-projection subspace is constructed by fitting a cubic polynomial through the data points known to be from that motion. Note that since the pair relation f_{ij} induces an ordering of dimensions (e.g. predicting ankle-angle from knee-angle is different than the converse) there are $N(N-1)$ possible pair predictors, twice as many as there are two dimensional subspaces.

The pair predictors may be viewed as each embodying one component of the constraint imposed by a movement. However, they really are static constructs in the sense that no temporal continuity is considered. Even without considering a full dynamic model of dance, it is possible to increase the reliability of the pair predictor $g_{ij}(t)$ by including a simple temporal filter:

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.

Finally, we need to combine pair predictors to represent all the constraints on a movement. To select which combination of predictors is best at representing a motion, we need to define the notion of predictor efficiency. Here we use the standard pattern recognition criteria of minimizing a weighted sum of false acceptances and false rejections. Our goal is to find a collection of pair predictors with the greatest efficiency:

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

Note that the approach to recognition described above defines an independent detector for each dance step; in principle these detectors can be executed in parallel. Thus we are *not* constructing a pattern recognition system because their is no constraint preventing the same input data as being recognized as two or more different dance steps.

The described quantities establish a method for representing and recognizing ballet steps. To apply this method we must be able to construct the necessary functions from the data.

4.4 Learning predictors.

We employ a supervised learning paradigm to determine the 2D-projection subspaces, the predictor curves, and thresholds to be used to represent each ballet step. The input data is annotated to indicate, for each time step, whether it is part of a particular dance step or part of a non-dancing pose. To find good identifiers, the system does a hierarchical search, successively constraining the region of phase space in which the movement lies.

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. This latter type is for classification [6]. 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 learning algorithm proceeds by first using the fitness function to select the acceptance threshold of a given pair predictor. A hierarchical exhaustive search is performed to choose the threshold that maximizes the fitness of the predictor for the current motion.

Next, the system determines the best k pair predictors. Since there is a reasonable number ($N(N-1)$) of pair predictors, the pairs are searched exhaustively, and the k best, as measured by the fitness function, are 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. In practice the false acceptances tend to be partially correlated, but ANDing predictions usually dramatically decreases the false acceptance rate while only slightly reducing the correct acceptances.

Note that 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. Instead, the system chooses from among the multiple representations the one which predicts best. In the case where multiple sets of Euler angles represent hip orientation, the best predictor tends to be Euler angles which are far from singularity and therefore less noisy.

The result of the learning process is a multidimensional tube-like volume in phase space, formed by intersecting the $N-1$ dimensional curved-hyperplanes associated with each pair predictor. This result raises the question: why construct the volume by compounding instead of directly optimizing all parameters simultaneously to find the best k-dimensional subspace space curve to represent the motion? 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.

5 Experimental Results

We have developed a recognition system based on the theory developed in the previous section. The input is Cartesian (XYZ) tracking data recorded from 14 points on the dancer's body at a rate of 60 frames/sec., which is converted to joint angles of six 1DOF and four 3DOF joints in the arms and legs (see figure 5).

The system was tested on a set of nine ballet steps which were chosen to be "atomic" in the sense that they are "pieces" or "syllables" of more complex ballet movements. Training and recognition data was used from two different dancers whose heights are 157 cm and 173 cm to test whether the system could do dancer-independent recognition. The input sequences were continuous, not segmented, so the system could make no implicit assumptions about where a dance step begins and ends. Finally, our tests were resubstitution estimates in which the same data is used for training and for testing; this choice was necessitated by an inadequate amount of test data.

The parameters used for learning and recognition were: right hip–x, right hip–y, right hip–z: an Euler angle set shown in Figure 4 in which the x–axis (front–back) is superior, the y–axis (side to side) is the middle axis, and the z–axis (along femur) is inferior; right hip–ϕ: part of another Euler angle set which measures deflection of the femur from vertical; right knee and ankle angles; and Z, the average height of the hips, scaled for the dancer's leg length.

As mentioned in the previous section, insensitivity to temporal variation implies that parameter derivatives should not be used as possible candidates for pair predictors. To simplify the learning task we did not include derivative dimensions in the phase space to be searched.

Figure 6 shows the raw data input to the learning and recognition system. Each column contains the data for one dancer executing 9 ballet moves in succession. The difference in duration (16 versus 22 seconds) show the degree of temporal variability in the data. The spikes in the data are missing data where the tracking system could not provide the three-dimensional position of some of the markers.

As developed, 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. 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.

We evaluate the results in two ways. First, we can consider each time step independently, analyzing how many false acceptances and false rejections each predictor generates. Second, we can count the number of *predictor errors* which looks at the behavior of the predictor with respect to the time interval that makes up the move. A predictor error occur if the predictor (1) outputs a detection interval during a different step or when no step is occurring; (2) outputs a detection interval more than once when there is only one occurrence of the step; or (3) fails to output a detection interval during a step.

Table 1 shows the number of timesteps falsely accepted and falsely rejected (FA and FR) with neither temporal smoothing nor ANDing. This can be viewed as a model-free pattern recognition approach to the problem where the system simply adjusts thresholds to minimize FA + FR. Many predictor errors occurred because the predictors often turned on or off for single timesteps, as can be seen in the last column. Note, however, that using just two variables, both the plié and relevé can be detected quite well, and in fact have no predictor errors.

Table 2 and the "announcement" pulses in Figure 7 show the reults of using compound predictors formed by ANDing three

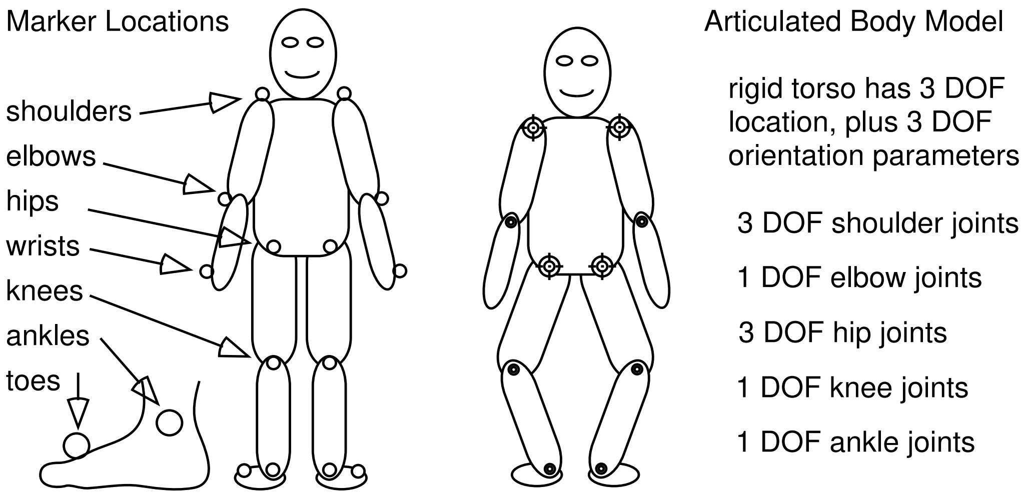

Figure 5: Tracking and human body models. The AOA system [3] returns XYZ data for 14 variables yielding 42 channels of data. From this data 24 model parameters are extracted: torso location and attitude – 6 parameters represented by X, Y, Z, yaw, pitch, and roll; shoulders and hips – 3 parameters each, represented by 3 Euler angles; elbows, knees and ankles – 1 parameter each represented by 1 angle.

Best pair predictors; No ANDing
cost ratio w = 1:1 temporal filtering time = 0

Move	Variables	K	$N-K$	FA (%)	FR (%)	Predictor errors
plié	hip–ϕ, Z	105	2175	0	1 (1)	0
relevé	hip–y, Z	161	2119	14 (0.7)	26 (9)	0
tendu	hip–y, Z	143	2137	47 (2.2)	37 (26)	7
dégagé	hip–ϕ, knee	97	2203	31 (1.4)	46 (50)	17
fondu	hip–x, knee	188	2092	29 (1.4)	88 (47)	19
frappé	hip–x, hip–y	175	2105	4 (.02)	166 (95)	6
développé	hip–ϕ, Z	298	1982	121 (6.1)	147 (50)	19
g. batt. side	hip–y, Z	108	2172	5 (.02)	79 (73)	6
g. batt. front	hip–x, hip–y	105	2175	17 (0.8)	15 (14)	6

Table 1: w is ratio of cost of FR (false rejection) to FA (false acceptance); N is total number of timesteps; K is number of timesteps during the move. This approach tries to maximize the number of timesteps properly classified while ignoring temporal continuity. False acceptance and false rejection receive equal penalties. The predictor error count [# of times the predictor turned on] − [# of occcurences of dance step] is high because the predictors often turn on very briefly during other movements.

Best for Annotation; 3–way ANDing
cost ratio w = 4.3:1 temporal filtering time = .3 sec

Move	Variables	K	$N-K$	FA (%)	FR (%)	Annot. errs
plié	hip–x, hip–ϕ, knee, Z	105	2175	0	1 (1)	0
relevé	hip–y, ankle, Z	161	2119	6 (.3)	21 (13)	0
tendu	hip–x, hip–y, hip–ϕ, Z	143	2137	0	31 (22)	0
dégagé	hip–x, hip–ϕ, knee	97	2203	0	15 (16)	0
fondu	hip–x, hip–y, hip–z, hip–ϕ, Z	188	2092	0	102 (54)	0
frappé	hip–x, hip–y, hip–ϕ, Z	175	2105	0	18 (10)	0
développé	hip–x, hip–z, hip–ϕ, Z	298	1982	52(2.6)	148 (50)	1
g. batt. side	hip–x, hip–ϕ, knee, Z	108	2172	0	1 (1)	0
g. batt. front	hip–x, hip–y, ankle, Z	105	2175	2 (.01)	32 (30)	0

Table 2: ANDing means that e.g. only timesteps accepted by all 3 plié detectors (individually temporally filtered) will be classified as plié. Each individual detector can allow higher acceptance rates because the conjunction lowers the probability of random false acceptance. Temporal filtering rejects shortlived acceptance periods. All this allows us to bias the individual predictors towards acceptance with a lower penalty on FA, and yet see very low rates of FA on the compound predictors.

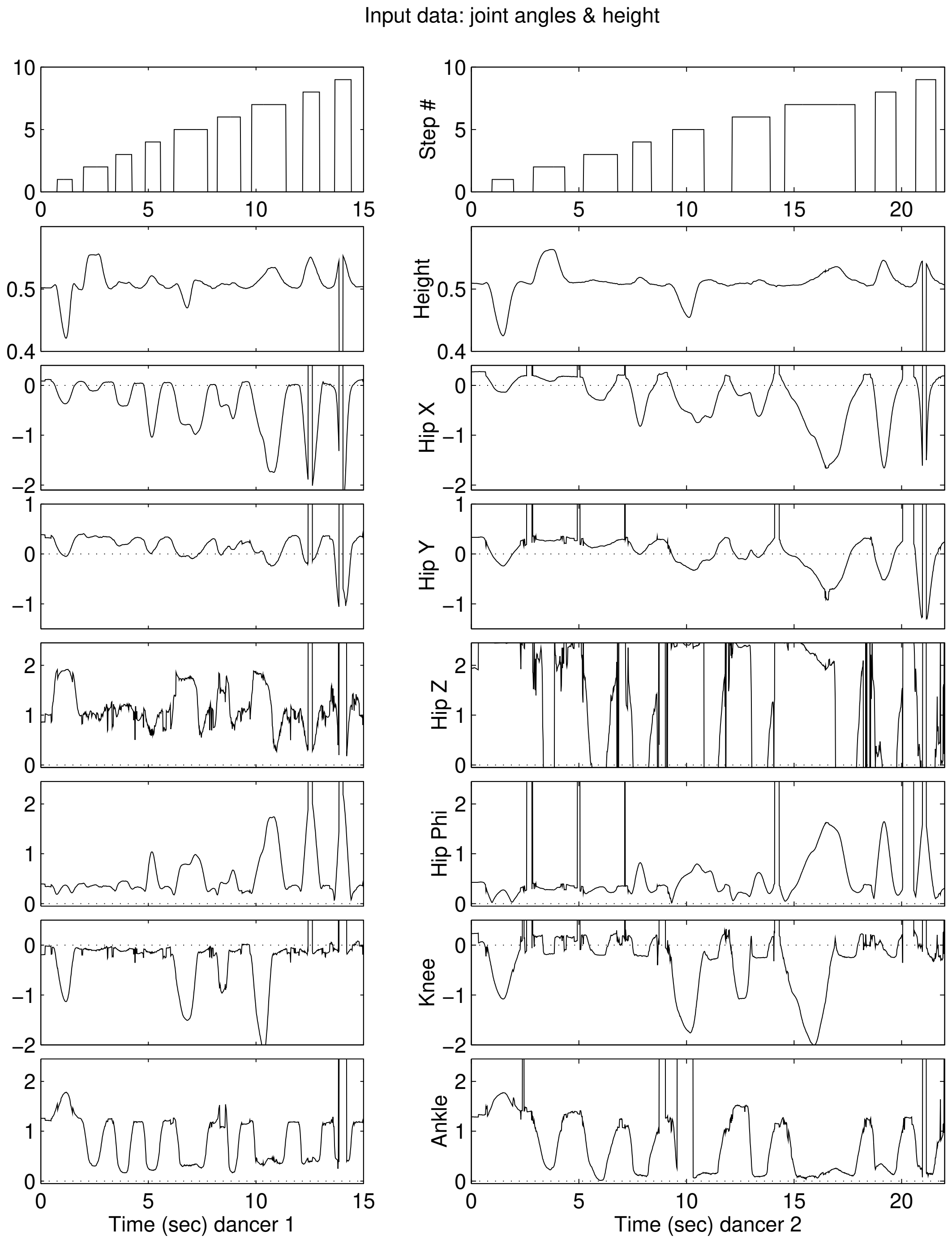

Figure 6: Input data for two dancers processed by dance step recognition system. The top graph is ground truth identification; the next is torso height scaled by dancer's leg length; three Euler angles representing the dancer's right hip; a fourth Euler angle from a different representation; and right 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 Hip ϕ) are missing data codes.

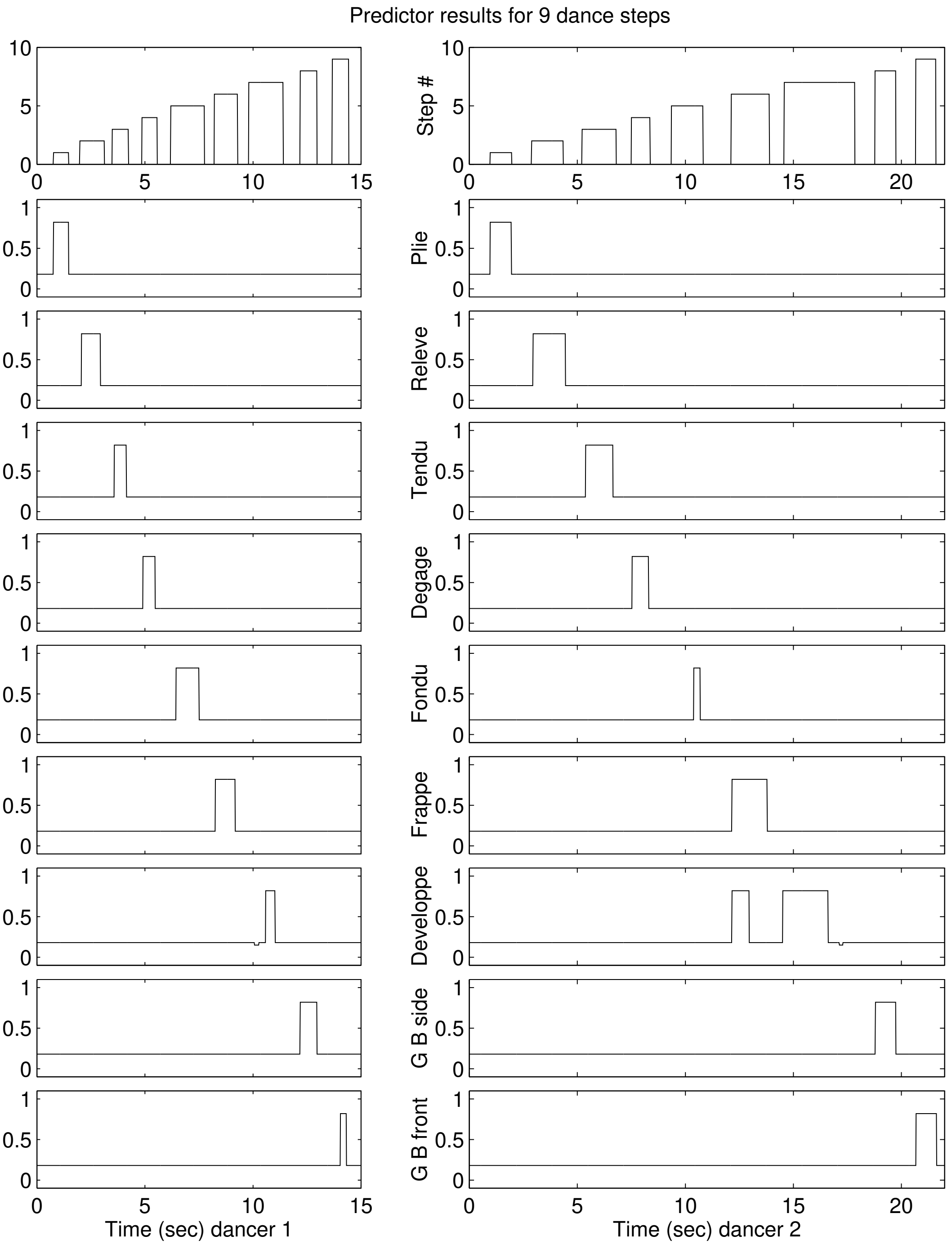

Figure 7: Predictor outputs for two sets of movements. Note the error for the développé predictor, which fires wrongly for frappé as well as correctly for développé on dancer 2.

pair predictors each of which has been temporally smoothed. As explained in section 4.4.1 a higher cost ratio w was used for the fitness function in anticipation of the filter effect of ANDing. Table 2 shows FA to be zero for most steps and FR to be acceptable, demonstrating the benefits of temporal smoothing and the compound prediction. As can be seen in figure 7, the system recognized all steps and produced only one error, viz. the développé predictor, which fires wrongly for frappé as well as correctly for développé on dancer 2. These movements have very similar beginnings: they both start by flexing the knee and hip such that the right foot rises parallel to the left calf. The system annotated all other steps correctly.

A demonstration of invariance to speed and extension can be seen in Figure 6. One dancer performs the sequence in 16 sec., the other in 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 correctly recognized these steps.

6 Summary

In this work we have developed a system for recognizing classical ballet steps from an input of Cartesian XYZ tracking data. Our examination of representation issues led us to a novel "phase space" representation borrowed from classical physics. Among the advantages of 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 performance. The learning method, which works automatically on movement sequences annotated with ground truth ballet step identifications, has two parameters which must be set manually: the penalty weight ratio for the fitness function which evaluates predictors, and the smoothing time constant.

The ballet step recognition system works extremely well on identifying ballet steps and moderately well on identifying their start and stop times. More work is needed to detect when a training movement is not "atomic" and if so, to automatically break it into atomic pieces. Also, the temporal continuity check could be augmented by learning bounds on velocities during training, and enforcing those bounds. It will be valuable to apply the system to other recognition tasks such as recognizing facial expressions.

Finally, we note that this representation is also suitable for some "natural," non ballet movements, specifically those that involve whole body movements. Though most natural 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 well modeled by a one degree of freedom space curve for the major body masses.

References

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

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

[3] Adaptive Optics Associates. *Multi-Trax User Manual*. Adaptive Optics Associates, Cambridge, MA, 1993.

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

[5] 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.

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

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

[8] Lee W Campbell. Recognizing classical ballet steps using phase space constraints. Master's thesis, Massachusetts Institute of Technology, Sept 1994.

[9] Claudette Cedras and Mubarak Shah. A survey of motion analysis from moving light displays. In *Proc. 1994 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 214–221. IEEE Press, 1994.

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

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

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

[13] 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.

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

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

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

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

[18] Thomas S. Huang and Arun N. Netravali. Motion and structure from feature correspondences: A review. *Proceedings of the IEEE*, 82(2):252–268, Feb 1994.

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

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

[21] 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.

[22] 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.

[23] M. P. Murray et al. Walking patterns of normal men. *J Bone and Joint Surgery*, 46A(2):335–360, 1964.

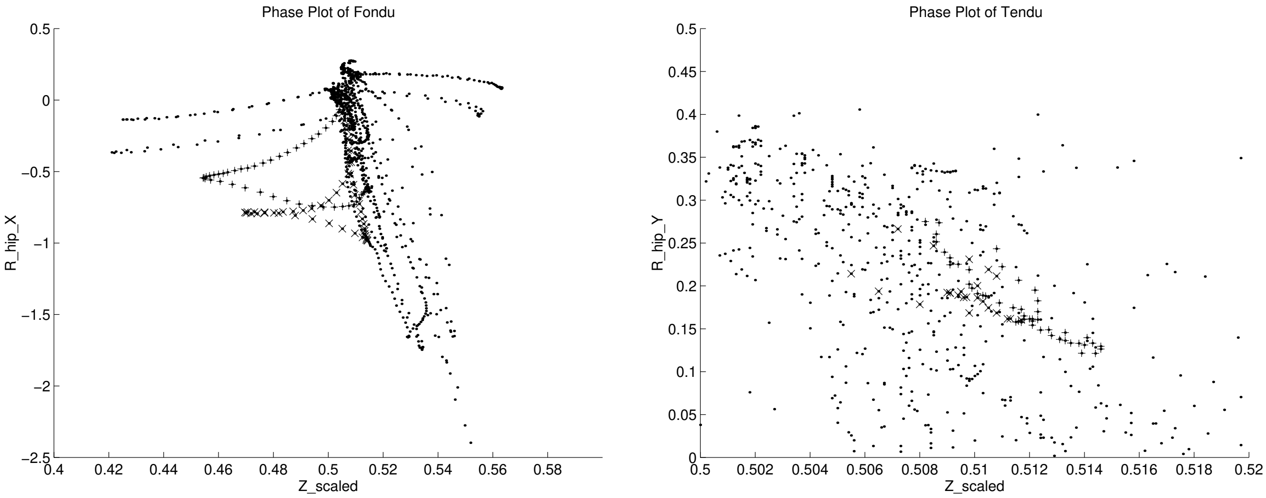

Figure 8: $\times$ and $+$ mark time steps during fondu and tendu for two dancers; "." marks time steps during other movements. The left figure illustrates curves that are loops rather than cubics, and offsets indicating differences in angle measurement due to differences in reflector placement. The right figure illustrates a spurious correlation. The change in both variables is small; the predictor learns a patch in phase space rather than a curve, and temporal smoothing eliminates other movements that pass briefly through the patch.

[24] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *IEEE Workshop on Motion of Nonrigid and Articulated Objects*, pages 64–69, Austin TX, Nov 1994.

[25] 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.

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

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

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

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

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

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

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

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

[34] 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.

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