ABSTRACT

Title of Document: A SENSORY-MOTOR LINGUISTIC FRAMEWORK FOR HUMAN ACTIVITY UNDERSTANDING

Gutemberg B. Guerra Filho, Ph.D., 2007

Directed By: Professor Yiannis Aloimonos, Department of Computer Science

We empirically discovered that the space of human actions has a linguistic structure. This is a sensory-motor space consisting of the evolution of joint angles of the human body in movement. The space of human activity has its own phonemes, morphemes, and sentences. We present a Human Activity Language (HAL) for symbolic non-arbitrary representation of sensory and motor information of human activity. This language was learned from large amounts of motion capture data.

Kinetology, the phonology of human movement, finds basic primitives for human motion (segmentation) and associates them with symbols (symbolization). This way, kinetology provides a symbolic representation for human movement that allows synthesis, analysis, and symbolic manipulation. We introduce a kinetological system and propose five basic principles on which such a system should be based: compactness, view-invariance, reproducibility, selectivity, and reconstructivity. We demonstrate the kinetological properties of our sensory-motor primitives. Further

evaluation is accomplished with experiments on compression and decompression of motion data.

The morphology of a human action relates to the inference of essential parts of movement (morpho-kinetology) and its structure (morpho-syntax). To learn morphemes and their structure, we present a grammatical inference methodology and introduce a parallel learning algorithm to induce a grammar system representing a single action. The algorithm infers components of the grammar system as a subset of essential actuators, a CFG grammar for the language of each component representing the motion pattern performed in a single actuator, and synchronization rules modeling coordination among actuators.

The syntax of human activities involves the construction of sentences using action morphemes. A sentence may range from a single action morpheme (nuclear syntax) to a sequence of sets of morphemes. A single morpheme is decomposed into analogs of lexical categories: nouns, adjectives, verbs, and adverbs. The sets of morphemes represent simultaneous actions (parallel syntax) and a sequence of movements is related to the concatenation of activities (sequential syntax).

We demonstrate this linguistic framework on real motion capture data from a large scale database containing around 200 different actions corresponding to English verbs associated with voluntary meaningful observable movement.

A SENSORY-MOTOR LINGUISTIC FRAMEWORK FOR HUMAN ACTIVITY UNDERSTANDING

By

Gutemberg B. Guerra Filho

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Yiannis Aloimonos, Chair
Dr. Cornelia Fermüller
Professor Larry Davis
Associate Professor David Jacobs
Associate Professor José Contreras-Vidal

Dedication

To Andreia with love.

Acknowledgements

When I came here, everything was strange to me. As many people, I came from far away and all the great differences I found made me miss home in a way bigger than distance. Today, I don't miss home as much as I did. Sure time works miracles, but the main reason I was able to overcome all challenges I faced during this journey was the good people I met in my path. Some of them gave me directions. Others walked with me along the way. And there are the ones who even carried me for some time. To all those people, who made me feel home, my deepest gratitude.

To God and my guard angel, for keeping me and my family safe and sound and for all the gifts and blessings I was presented with. I can only hope that I will serve well to Your purpose.

To my wife, Andreia, for every single day of poetry in my life. For making cold days warmer, sunsets more beautiful, Sunday mornings complete, and the future worth of dreaming about.

To my parents, Gutemberg and Rosa, for giving me good values, the roots of my character, all the structure I needed to grow, the memories of a great childhood, and a past worth of remembering.

To my advisors, Yiannis and Cornelia, for the warm environment they created in the Computer Vision Lab. For the great advices, ideas, motivation, patience, mentoring, and for making me address the right questions with enthusiasm and excitement.

To my family, for the support and encouragement that only the close ones know how to give. To my friends, Henrique Andrade, Renato Ferreira, Indrajit Bhattacharya,

Zoran Majkic, Michael Beynon, Camilo, and Maximiliano Guimarães, for providing all help and guidance that one needs in a new place.

To my colleagues, Abhijit Ogale, Patrick Baker, Jan Neumann, and Alap Karapurkar, for the enlightened discussions, brainstorms, and specially for sharing the same academic boat with me for a little while.

To Dr. Hanan Samet and Dr. Larry Davis, for their support in my academic decisions.

To the additional members of the committee, Dr. David Jacobs and Dr. José Contreras-Vidal, for their valuable suggestions and comments. The support of CAPES, NSF, and DARPA is gratefully acknowledged.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Activity understanding is an important component of human intelligence. Natural intelligent systems perceive events occurring in the environment, reason about what is happening, and act accordingly. This process involves mapping observed and generated motor sequences onto a vocabulary of actions. This vocabulary represents sensory-motor patterns learned previously and stored according to some knowledge representation.

An artificial cognitive system with commensurate abilities may require a symbolic structure for reasoning about human activities. However, the semantic interpretation of a symbolic representation system, such as natural language, cannot be based only on meaningless arbitrary symbols. The *symbol grounding problem* [Harnard, 1990] addresses this semantic gap and suggests that the primitives of a formal symbolic system should be associated with grounded representations connected to physical experience in the world.

A *grounded representation* is a sensory-motor projection of objects, actions, and events to which elementary symbols refer. With regards to events associated with human activities, a *sensory-motor projection* consists in the mapping from a non-symbolic analog representation of human activities in the world to a non-arbitrary symbolic representation according to invariant features, which allow cognitive tasks.

One important aspect of artificial cognitive systems is the need for computers to be able to share a conceptual system with humans. Concepts are the elementary units of reason and linguistic meaning. Many researchers hold the philosophical position that all concepts are symbolic and abstract and therefore should be implemented outside

the sensory-motor system. This way, meaning for a concept amounts to the content of a symbolic expression, a definition of the concept in a logical calculus.

An alternative approach states that concepts are grounded in sensory-motor representations. This sensory-motor intelligence considers sensors and motors in the shaping of the hidden cognitive mechanisms and knowledge incorporation. There exists a variety of studies in many disciplines—such as neurophysiology, psychophysics, and cognitive linguistics—suggesting that the human sensory-motor system is indeed deeply involved in concept representation.

Knowledge of actions is crucial to our survival. Hence, human infants begin to acquire actions by watching and imitating the actions performed by others. With time, they learn to combine and chain simple actions to form more complex actions. This process is analogous to speech, where we combine simple constituents called phonemes into words, and words into clauses and sentences. Humans can recognize as well as generate both actions and speech. In fact, the binding between the cognitive and generative aspects of actions is revealed at the neural level in the monkey brain by the presence of *mirror neuron networks*, i.e., neuron assemblies which are activated when the individual observes a goal-oriented action (like grasping) and also when the individual performs the same action. All these observations lead us to a simple hypothesis: Actions are effectively characterized by a language. This is a language with its own building blocks (phonemes), its own words (lexicon), and its own syntax.

The realm of human actions may be represented in at least three spaces: sensory space, motor space, and natural language space. Therefore, we can imagine that

actions possess at least three languages: a sensory language, a motor language, and a natural language as Figure 1.1 shows. The sensory language lets us perceive actions, the motor language lets us produce actions, and the natural language lets us communicate about actions. The sensory domain covers the form of human actions when perceived. The motor domain covers the underlying control sequences that lead to the generation of movements. The linguistic domain covers symbolic descriptions of natural actions. We took the hierarchical structure of natural language (e.g., phonology, morphology, syntax) as a framework for structuring not only the linguistic system that describes actions, but also the sensory and motor systems. We defined and computationally modeled sensory-motor structures that are analogous to basic linguistic counterparts: phonemes (the alphabet), morphemes (the dictionary), and syntax (the rules of combination of entries in the dictionary) using data-driven techniques grounded in actual human movement data.

We study a language that maps to the lower-level sensory and motor languages and to the higher-level natural language. By modeling actions as a language in each space, we can formulate many interesting problems as language translation problems that convert representations from one space to another: (a) video annotation for creating text descriptions of activity from a video, (b) natural-language-driven character animation, (c) training robots by imitation using video, and (d) controlling robots with natural language.

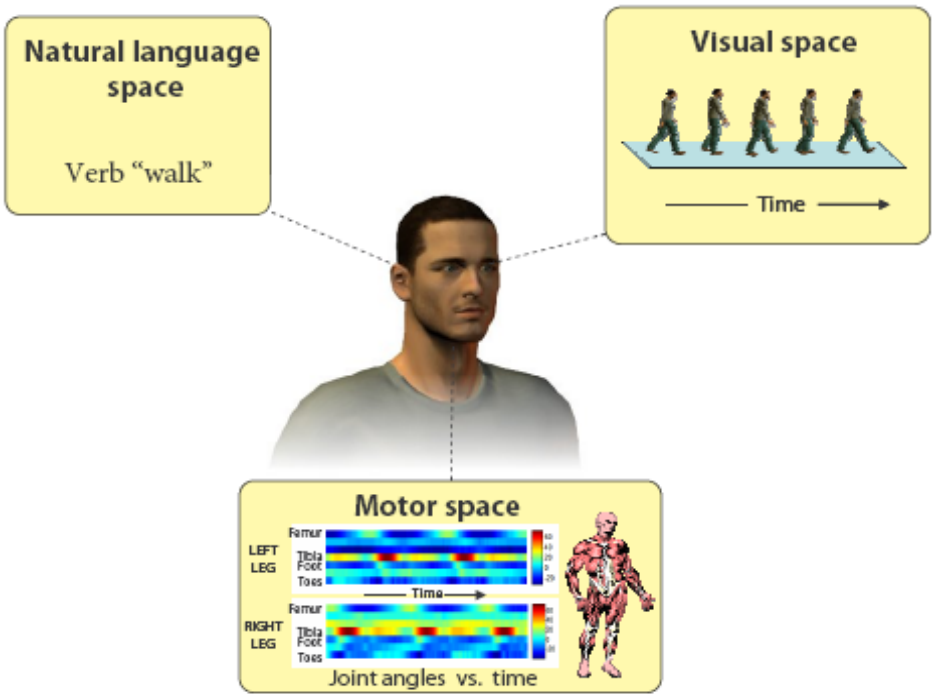


Figure 1.1. Three language spaces for human action (courtesy of Abhijit Ogale).

In the sensory pathway, the cognitive understanding involves the analysis (parsing) of observed action sequences towards an organized *praxicon*, a structured lexicon of human actions, movements, or praxis, previously learned and stored. In the motor pathway, the cognitive process concerns the synthesis (generation) of executed action sequences from this praxicon.

A sensory-motor praxicon is organized according to some knowledge representation. In this dissertation, we advocate a linguistic representation to support artificial cognitive systems. The design of such an artificial cognitive system involves the construction of a repertoire or vocabulary of activities based on sensory-motor representations. A large variety of behaviors in this set increases the chances that an observed action is recognized and the possibilities of motor strategies in achieving some goal.

Scope and Contributions

This dissertation establishes a very fundamental result, namely that the space of human actions is characterized by a language that can be learned from human motion measurements using modern and novel techniques.

Artificial cognitive systems could become more powerful if they possessed models of human actions. Thus, we investigate the involvement of sensory-motor intelligence in concept description and, more specifically, the structure in the space of human actions. In the sensory-motor intelligence domain, our scope centers on the representation level of human activity. We contribute to the modeling of human actions with a sensory-motor linguistic framework. An artificial cognitive system with sensory-motor representations can learn skills through imitation, better interact with humans, and understand human activities. This understanding includes reasoning and the association of meaning to concrete concepts.

Closing this semantic gap involves the grounding of concepts on the sensory-motor information. In this dissertation, we contribute to the grounding of concrete concepts by modeling human actions with a sensory-motor linguistic framework. The grounding process starts from sensory data (e.g., video, audio), where objects are detected, recognized, categorized, and identified. At this level, human body parts might become features extracted from visual input and, consequently, allow the 3D capture of human movement.

We seek information on human actions that correspond to general concrete human action. *Concrete human activity* concerns observable, voluntary, and meaningful movement. We concentrate in the motor domain and we exclude thinking and feeling

activities. A non-exhaustive list of concrete activities includes breathing patterns, eye movement, facial expressions, head movement, postures, orientation in space, change of stance, trunk movement, limb action, manipulation, and locomotion.

The problem addressed in this dissertation is to learn representations for human activity. Motion capture data is processed towards the discovery of structure in this space. This input contains the essential 3D specification of human movement necessary to the mapping towards visual and motor spaces. We hypothesize that there exists a language—in a formal sense—that describes all human action, and then show how we could obtain this language using empirical data.

We concentrate on the modeling of sensory and motor information in a higher-level than perception and generation. Perception, the mapping from images to cognitive representations, involves detection, recognition, categorization, and identification. Generation, the mapping from cognitive representations to motor control, involves planning, control, retargeting, and dynamic stability. Recent results in these areas show the feasibility of mapping joint angle data into actuator control [Huang at al., 2001; Matsui at al., 2005; Pollard at al., 2002; Ude at al., 2000].

We propose a linguistic framework for the modeling and learning of human activity representations. The linguistic framework is used to represent human movement with a grounded symbolic system. By grounded we mean that symbols have a non-arbitrary mapping to the sensory-motor primitives. We seek to provide a flexible representation, proposed here as a *Human Activity Language* (HAL), to model the sequential and parallel aspects of human movement. Our sensory-motor language allows perception and generation of hundreds of human actions modeled in a compact

structure. This structure—organized in terms of kinetology, morphology, and syntax—has the flexibility required to handle numerous behaviors using the parsing and generation aspects of a language.

Kinetology, the phonology of human movement, finds basic primitives for human motion (segmentation) and associates them with symbols (symbolization). A kinetological system transforms continuous motion signal into a non-arbitrary discrete representation of human movement. Kinetology provides a symbolic representation that allows synthesis, analysis, and symbolic manipulation of human movement. We introduce five principles on which kinetology should be based and evaluated: compactness, view-invariance, reproducibility, selectivity, and reconstructivity. These properties provide a reasonable way to evaluate a kinetological system. Besides providing the foundations to the inference of more structure in human movement, kinetology also has applications to compression, decompression, and indexing of motion data [Guerra-Filho and Aloimonos, 2006b; Kovar and Gleicher 2004]. We present experiments on compression and decompression of motion data. These experiments demonstrate the parsing and generation of movement in the lowest level.

The *morphology* of a human action relates to the essential parts of the movement and its structure. A single action *morpheme* represents the least amount of movement with a purposeful goal, i.e., meaning. The morphology of a specific human activity consists of the set of actuators involved in the activity, the synchronization among these actuators, and the motion pattern associated with each participating actuator. We propose a novel formal grammar system as an action morpheme, where each

component grammar corresponds to an actuator. To learn action morphemes, we present a grammatical inference methodology and introduce a heuristic parallel learning algorithm to induce a grammar system representing a single action. In *morpho-kinetology*, a praxicon is empirically built by inducing grammar systems for all actions in a large motion capture database. *Morpho-syntax* amounts to explore the morphological organization of a praxicon towards the discovery of more structure in a Human Activity Language.

The results of our approach are both theoretical, concerning the heuristic inference of a parallel grammar system, and empirical, in terms of human movement learning and representation. An advantage of parallel learning over plain sequential learning is that problems with overgeneralization are resolved in parallel learning. Sequential learning is able to infer the structure of a single sequence of symbols. This structure corresponds to a forest of binary trees, where each node in a tree is associated with a grammar rule in a normal form. A sequential learning algorithm may keep merging adjacent root nodes into single rules (trees) and, consequently, overgeneralization happens when unrelated rules are combined and generalized. This happens mostly in higher-levels of the grammar tree. In parallel learning, we consider all joint angles simultaneously and we use the learned synchronized rules to resolve overgeneralization. Nodes are merged only if the new rule is synchronized with other rules in different components of the grammar system. This way, overgeneralization is avoided since synchronization guarantees a relationship between the merged rules.

The *syntax* of human activities involves the construction of sentences using action morphemes. A sentence may range from a single action morpheme (nuclear syntax)

to a sequence of simultaneous sets of morphemes. The *nuclear syntax* consists of a noun phrase and a verbal phrase. A noun phrase in a sentence corresponds to the active joints (noun) modified by an initial posture (adjective). A verbal phrase includes the changes each active joint experiences during the activity execution (verb) and a point in a reduced dimensional space (adverb) which serves to modify the activity. The analogy to lexical categories (nouns, adjectives, verbs, and adverbs) gives more intuition about the human activity model. These intuitive categories suggest a simple representation that may be used to interface between users and artificial cognitive systems. This analogous set of lexical categories lacks prepositions and conjunctions that are probably related to preparatory movements performed between actions.

Nuclear syntax, especially adverbs [Rose at al., 1998], relates to the motion interpolation problem. Motion interpolation or morphing adapts an exemplar motion to new circumstances. Interpolations involve parameterized spaces where high-level motion properties (target location, locomotion style) are represented as interpolation weights. Parametric synthesis allows accurate generation of any motion for an entire space of motions, parameterized by continuously valued parameters.

We introduce a novel representation for motion adverbs that addresses the interpolation problem. This representation solves the semantic problem (intuitive mapping between external and internal motion representations) and the universality problem (a single adverb representation is used for all actions). The modeling of adverbs involved a simple quadratic interpolation that resulted in adverbial components for each action.

Besides the nuclear syntax, parallel and sequential syntax are required to handle the simultaneity and sequencing of actions, respectively. The sets of morphemes represent simultaneous actions (*parallel syntax*) and a sequence of sets of morphemes models the concatenation of activities (*sequential syntax*).

Parallel syntax relates to the *splicing problem* that concerns the combination of motions of different body parts. The general splicing problem involves transferring any subset of the body from one motion to another. Most splicing approaches only address the details of how to merge the motions smoothly [Heck at al., 2006]. They assume the sets of body parts being merged are known. Our method actually learns the set of essential actuators for each action such that motion splicing considers only these sets when blending movements of different body parts and different actions. Parallel syntax introduced constraints to the splicing of human actions. This way, our framework generalizes whole-body techniques in many aspects. Even our motion interpolation method considers only the essential actuators of each action.

Sequential syntax is proposed as an alternative method for the *transitioning problem*. A transition is a segment of motion that seamlessly attaches two motions to form a single longer motion. Therefore, the transitioning problem concerns the motion concatenation towards longer sequences of motion. Transitions between different actions are found according to the structure of morphological grammars. Basically, a transition is feasible if the actions share primitives in their respective motion patterns. In what concerns transitioning, our Human Activity Language is more compact, efficient, and structured than state-of-the-art approaches such as motion graphs and its variants [Arikan and Forsythe, 2002; Kovar at al., 2002; Lee at al., 2002].

The experimental validation of our linguistic framework is performed on real human motion obtained by a motion capture system. Our motion-capture database contains around 200 different actions corresponding to concrete English verbs associated with observable voluntary meaningful movement (see Appendix A). The actions are not limited to any specific domain. Instead, the database includes actions of several types: manipulative (prehension and dexterity), nonlocomotor, locomotor, and interaction.

Sensory-Motor Embodiment

To consider the tasks related to activity understanding, we introduce a sensory-motor system model with four subsystems: perception, recognition, motor planning, and action. In this model, a sensory-motor language plays a central role in supporting activity understanding as a common representation for sensory and motor information, as shown in Figure 1.2.

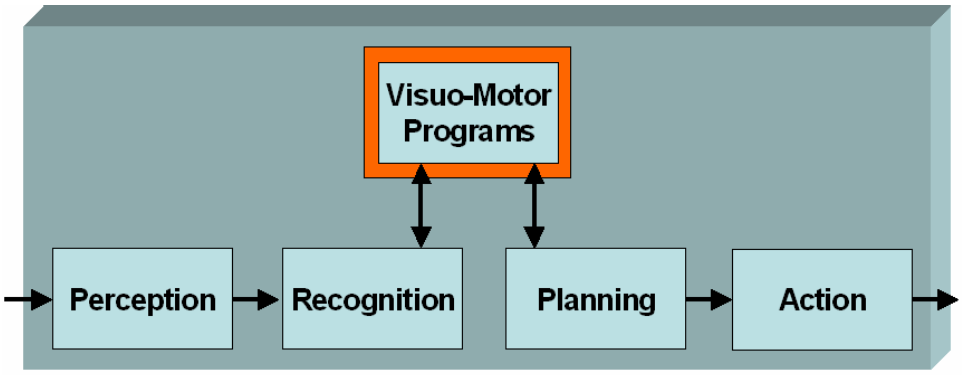


Figure 1.2. A sensory-motor system model.

A perception subsystem takes the sensory input and extracts higher-level representations for human actions. These representations are parsed and possibly matched to sensory-motor programs by a recognition process. If the action vocabulary does not contain the observed action, no matching is found and learning may occur

through imitation. The imitation process searches for a physically feasible plan to execute the observed unknown action in the action subsystem.

Sensory and motor representations for human activity are coupled by embodiment. These two aspects of human activity are abstracted to a common ground through *embodiment* which is, ultimately, the consideration of the human body into the modeling process. In this dissertation, we focus in the discovery of a common embodied symbolic language.

Visual Representations

Vision detects whatever alters the pattern of light reaching the eye. There are two kinds of receptor cells in the retina (the end organ of vision): one for seeing fine detail and another for movement [Clark, 1963: 274]. The visual system is an organization that maps into a simple grammar: something (seen by foveal vision) moves (seen by the rod cells in the retina). This organization is the lowest level of representation for visual perception. In higher levels, the visual representations may range from motion fields to 2D joint angles derived from a stick model of the human body, as shown in Figure 1.3.

Global representations are the lowest level of visual representation which captures the whole body motion. *Structured representations* are higher-level representations that may record only the motion of specific structural components of the human body [Boyd and Little, 1997]. A structured representation requires tracking of specific body parts (e.g., joints) of the actor while simplifies the classification process involved in movement recognition.

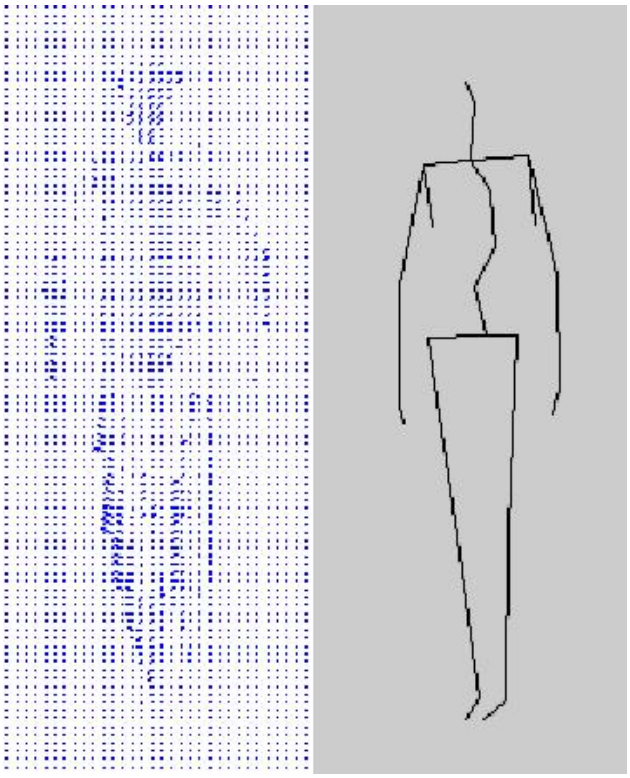


Figure 1.3. Visual representations from motion field to stick model.

The design of visual representations and the transformations between them are denoted as the *visual perception process*. The abstraction in the visual perception process starts with global representations towards finding more structured representations until there is enough embodiment to interact with the motor domain.

The abstraction process transforms global representations into structured representations. We suggest that this abstraction process consists in a gradual transformation of representations with an increasing level of embodiment. At each step, the process extracts a set of features using an embodied constraint. An embodied constraint is applied to a representation to get more structure into a higher-level representation. One example of an embodied constraint considers the movement of points in the same body part as rigid to detect different articulated body parts. A more

abstract constraint may use the topological connections of body parts to identify each detected body part with a segment in a human body model.

When the highest level representation (e.g., human stick model) is reached, features of joints are extracted. Some features of joints are angular position, angular velocity, and angular acceleration. Features from each frame are treated as time-varying scalar values and instants which are at the maxima or minima are view-invariant. This suggests a mapping from 2D to 3D features and, ultimately, a feasible way to map from a high-level visual representation to motor primitives.

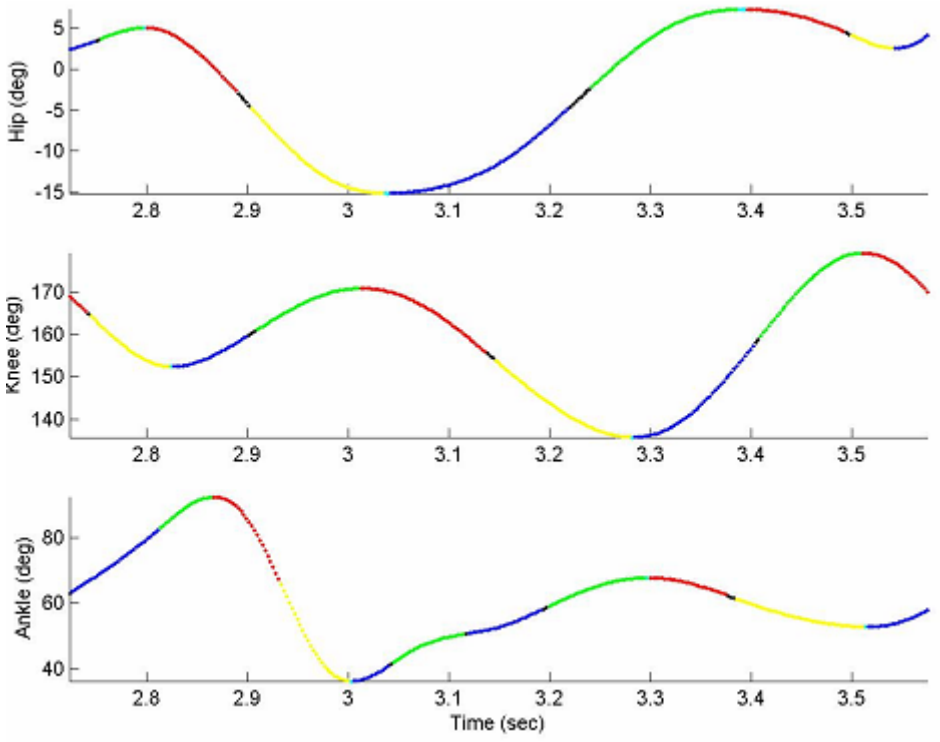


Figure 1.4. Joint angle functions for ankle, knee, and hip during jog activity.

One instance of the visual perception process is achieved by a motion capture system. We captured videos featuring 90 different human activities and the corresponding three-dimensional reconstruction for trajectories of body parts was found using our

own optical motion capture system [Guerra-Filho, 2005]. Given this three-dimensional reconstruction, joint angles were computed to describe human movement, as shown in Figure 1.4.

Motor Representations

Muscles are stimulated by electrical impulses (action potentials) that travel from a nerve to a muscle. The nerve is activated when a threshold current is achieved and it transmits a single packet of electric charge at a time. Each nerve action potential activates the muscle propagating another action potential into the muscle fibers to cause contraction. A single action potential only activates the muscle fibers for about 0.002 second (single twitch). To perform longer smooth controlled muscular contractions, the muscle needs to be stimulated repeatedly. The brain will send a stream of impulses through a certain number of nerves to the muscle to activate a proportional number of fibers so the muscle can contract and the corresponding force required is achieved.

All basic moves a human body can perform result from single muscle activations. The activation of muscles on the skeleton (mechanical behavior) is usually modeled by a number of force vectors. A *motor state* of the human body at a particular time is represented by a set of values, where each value corresponds to the force exerted by a certain muscle. Since the number of fibers activated in a muscle is discrete, each force has a discrete number of possible activation levels. These levels are the most fundamental units a human being can use to construct more complex actions and compose an alphabet of muscle activations. The motor state is an initial representation for human actions.

Although internal motor activity is constantly occurring, movement is any external observable motion. Different muscles collaborate to perform some specific anatomic action on a particular body part. An *anatomic action* corresponds to a resultant force for a system of force vectors associated with some muscles and, usually, acting on the same body part. Anatomic actions are the most basic movements that are visible and, hence, they are a starting point for the cyclic cognitive process between visual and motor representations.

The individual muscle activations are simultaneous while producing anatomic actions. In general, anatomic actions can be divided into flexion-extension (bending-straightening movement), abduction-adduction, and rotations. Most movement patterns are a combination of these muscle movements. An anatomic action performed by a specific joint and occurring in a particular anatomic plane (e.g., traverse, frontal, sagital) corresponds to a degree of freedom (DOF) in a human body.

An anatomic action corresponds to a subset of the motor state. For example, the "elbow extends" action corresponds only to the activation of the following muscles: anconeus, brachioradialis, and triceps brachii.

Each anatomic action corresponds to a resultant force for a system of force vectors associated with some muscles and, usually, acting on the same body part. Using an alphabet to represent scalar values associated with force resultants, an *anatomic state* is a set of symbols representing the resultant forces associated with each particular anatomic action. In this dissertation, we use another approach to action definition. A joint angle time-varying function for all DOFs represents a human action. We use this

initial representation in the derivation of a grounded symbolic representation: a Human Activity Language.

Areas of Application

The availability of a language characterizing human action has implications with regards to the grounding problem, to the universal grammar theory, and to the origin of human language and its acquisition process. Besides these theoretical issues, a linguistic representation for human activity has several practical advantages. A compact specification for human activity leads to compression and better efficiency. Once a symbolic linguistic representation is provided, the natural language processing and speech recognition fields are sources of methods that could be applied to activity understanding. A non-arbitrary symbolic representation allows the use of techniques of symbolic reasoning for inference and other cognitive tasks (e.g., recognition) on human activities. This framework could also be used as a basic module of a symbolic query language for the processing of multimedia data.

A language for human activity involves several challenging problems and has impacts and applications in many areas. A symbolic representation for human activity materializes the concept of motor programs and enables the identification of common motor subprograms used in different activities. This way, the discovery of such language allows exploring how a motor activity vocabulary is organized in terms of its subprograms. An evidence of common motor subprograms is the theory of motor tapes. *Motor tapes* [Hoyle, 1983] are explicit representations of a movement trajectory in memory. When an agent needs information on how to perform an action, it finds the appropriate template in memory and executes it.

In kinesiology, athletic performance analysis optimizes the training process and improves performance. In biomechanics, rehabilitation medicine detects, describes anomalies, and helps in the development of treatments. In performing arts, motion representations interface with dance notation systems.

Humanoid robots are designed to interact with humans and to assist them in several tasks. To be more effective towards a seemingly interaction, a similar appearance to humans is an important requirement. Besides similar appearance, robot behavior should be as natural as possible. Consequently, perception and generation of human activities by a humanoid robot should be included in its artificial cognitive system. The integration of analysis and synthesis of movement in this system leads to an easier way of programming motor skills in humanoid robots: learning through imitation [Atkeson and Schaal, 1997; Schaal, 1999]. This integration is implemented as the representation and modeling of human movement in the cognitive system.

In robotics, adequate movement models are detailed domain knowledge of the solution for complex nonlinear dynamics problems related to motor coordination. The representations make these problems highly structured and suited for path planning and trajectory tracking of motor control. A sensory-motor Human Activity Language assists humanoid robots to generalize the planning and control of motor activities while using a vocabulary of human actions.

In computer graphics, data-driven or example-based computer animation uses motion capture data to automatically generate realistic motion for virtual characters. In this context, one of the main challenges of animation is to reduce user interaction through

the reuse of motion data. Automation should also include flexibility towards novel movements while preserving the realism found in original motion.

A step towards automatic animation requires a system that is able to generate as many actions as possible. This task involves the use of human activity models and the structure of a large praxicon. This vocabulary assists a user in the specification of what the virtual character is supposed to do. This way, we suggest the use of a collection of real motion data that resembles more closely a vocabulary of activities. We propose a linguistic approach to model and construct such a praxicon. This approach is able to integrate several motion synthesis problems related to data-driven computer animation in a single unifying framework. In this dissertation, we discuss motion interpolation, splicing, and transitioning. However, we plan as future work to consider other problems such as retargeting, motion editing, and style generation. We intend to investigate the placement of these problems in our Human Activity Language. Further, each motion synthesis problem corresponds to an inverse motion analysis problem. Therefore, besides a generative aspect, this approach also supports motion analysis where human movement is parsed when facing an action cognition task.

In computer vision, surveillance is achieved with automatic activity detection and recognition based on action representations. They also assist video annotation with efficient storage, transmission, editing, browsing, indexing, and retrieval of the motion data in visual media. Basically, low-level features in the visual data are mapped explicitly or implicitly into higher-level features representing human movement. These features are parsed according to our linguistic framework and,

consequently, concrete reasoning is performed on this grounded linguistic space. We believe multimedia applications will ultimately include all types of sensory data. Current applications involve mostly visual and audio information. However, the integration of further sensory data and motor information is extremely relevant.

Human-centered computing involves conforming computer technology to humans while naturally achieving human-machine interaction. In a human-centered system, the interaction focuses on human requirements, capabilities, and limitations. These anthropocentric systems also focus on the consideration of human sensory-motor skills in a wide range of activities so that the interface between artificial agents and human users accounts for perception and action in a novel interaction paradigm. This leads to behavior understanding through cognitive models that allow content description and, ultimately, the integration of real and virtual worlds.

Chapter 2: Related Work

In this chapter, we review related work to human activity understanding. The subjects discussed are the inspiration for a sensory-motor approach, the motivation for a linguistic framework, symbolic representations of human actions, motor primitives, motion data compression, learning through imitation and movement segmentation, the semantic gap bridging, automatic computer animation, markerless motion capture, action recognition, grammatical inference, and grammar systems.

The modeling of human movement into sensory-motor representations has been studied in many fields such as computer graphics [Ilg at al., 2004; Mezger at al., 2005], computer vision [Del Vecchio at al., 2003], robotics [Billard and Matarić, 2001; Matarić, 2002; Schaal, 1999], and neuroscience [Ahmed at al., 2002; Caelli at al., 2001; Etou at al., 2004; Mori and Uehara, 2001; Nakazawa at al., 2002].

Sensory-Motor Inspiration

Sensory and motor processes such as perception and action are fundamentally inseparable in cognition [Varela at al., 1991: 173]. The Broca's region in the human brain is related to various functions ranging from perception to action [Nishitani at al., 2005]. This perception-action link in Broca's area involves learning (e.g., language and skill acquisition) through imitation.

Mirror neurons are brain cells which activate when a monkey performs a specific action with its hand [Gallese at al., 1996]. The same neurons also fire when the monkey observes the same action. The mirror neurons in Broca's region were not activated when human subjects watched an action that is not in the observer's motor

vocabulary [Buccino at al., 2004]. This evidence suggests that action recognition is another function related to Broca's area.

The functionality of Broca's region in the brain and the mirror neurons theory suggest that perception and action share the same knowledge structure that provides common ground for sensory-motor tasks such as recognition and motor planning along with higher-level activities.

We consider this common sensory-motor representation to be at an imagination or visualization level of an artificial cognitive system, where simulation tasks, such as computer animation, and preparation for lower-level cognitive tasks are performed. The lower-level tasks are concerned with proper visual perception (e.g., motion capture from images) and with actual motor generation (e.g., computation of torque at joints).

Higher-level tasks involving logic reasoning and natural language may also be grounded on this common sensory-motor representation. Some research shows that language is semantically grounded on the motor system [Glenberg and Kaschak, 2002], which implies the possibility of a linguistic framework for a grounded representation. A linguistic representation can give rise to a higher-level specification of motion by using compositional and recursive structures.

Linguistic Motivation

Inspiration for a linguistic approach comes from converging evidence in several fields of science. Similarly to spoken language, movement patterns are composed of elements in combination and sequences, but they may not be organized exactly like language because dimensions are qualitatively different. Speech can be characterized

as interleaved patterns of movements, coordinated across articulators [Armstrong at al., 1995; Studdert-Kennedy, 1987].

Humans make finely controlled movements that produce invisible (or barely visible) but audible gestures in the throat and mouth. The information about these movements is broadcast to the environment for the purpose of communication. Human movement (action, activity, or behavior) also has a communication aspect. Any time a subject acts, his/her image is optically broadcasted and communicates the intentions and other information associated with the action. Otherwise, there would be no need to deliberately keep ourselves in places where we cannot be seen by others (i.e., hide) when we want to avoid awareness of our presence. Since the same general model can describe both spoken and signed languages; we believe language is based in human body movement, which is a materialization of a more fundamental model or representation framework.

The description of acoustic and optic gestures uses the vocabulary of neuromuscular activity as a generalization of the vocal tract grammars at phonological level. Visible gesture words or sentences could have provided the behavioral building blocks associated with neuronal group structures for constructing syntax incrementally, behaviorally, and neurologically [Edelman, 1992].

Observations of people with brain injuries and diseases, coupled with dissection of their brains has shown that areas of anterior and parietal cortex in the left hemisphere of the cerebrum provides control for both vocal and manual activity including the hierarchical organization of manual object combination, signing, and speech [Greenfield, 1991; Kimura, 1981; Poizner at al., 1987]. The functionality of Broca's

region in the human brain also relates to language tasks [Nishitani at al., 2005]. The evidence of such a region in the brain with language and action functions is another inspiration for a linguistic approach to the representation of human activities in an artificial cognitive system.

Spoken language and visible movement use a similar cognitive substrate based on the embodiment of grammatical processing. For example, during walking acquisition, the human infant follows a developmental sequence that is not dissimilar from the sequence followed in language acquisition. The regularity of this developmental sequence is due to more basic underlying bio-behavioral forces.

Stages in motor development reflect neuromuscular maturation. The fundamental stages of sign language and spoken language acquisition are the same [Volterra and Erting, 1990: 302-303]. Infants go through a babbling stage, in which they manipulate the sublexical elements [Petitto and Marentette, 1991]. Language develops through social interaction since a word meaning is learned when heard or seen used by someone else in a context that made the relation between word and meaning reasonably unambiguous. Once language is acquired at a sufficient level, the meaning of unfamiliar words is determined by linguistic inference from its context.

Body movements are linked to visual perception, to recognition, to perceptual categorization (words must be sorted into categories: nouns and verbs), to memory, to learning, to concept formation, to primary self-consciousness and to consciousness of others, to pre-syntax, to language, to thinking, and to higher-order consciousness [Edelman, 1989: chapter 6].

Linguistics Foundations

Usually, linguistic is what we can write down stripping its emotional content and communicative intent [McNeill, 1985: 351]. A *language* consists of a system for making words, a system for making sentences out of words, and a system for reconciling conflicts between the first two [Lecture, "What is language," by Lyons at Christ's College, Cambridge 1977].

Phonology

Phonology is the system that selects certain speech sounds from all possible speech sounds and presents them as phonemes, the segments composing words. Different languages select for use different phoneme classes from among all the possible sounds that the human vocal tract is capable of emitting.

Usually, a phonetic description consists of a linear sequence of static physical measures, either articulatory configurations or acoustic parameters. Another approach characterizes phonetic structure as patterns of articulatory movements. A phonetic representation is a characterization of how a physical system changes over time [Browman and Goldstein, 1985: 35]. Muscular activity produces movement and gestures by moving the articulators along a trajectory. In a model of speech, words may be complexes of muscular activity temporally ordered, but not in the serial segmental way as in classical linguistic theory [Mowrey and Pagliuca, 1995].

Speech can be segmented into a linear stream of phones, which are analyzed into sets of features and abstracted as phonemes. However, the organization of human movement is simultaneous rather than sequential. Even though, sequentiality matters

at all levels of description since articulators must also follow a certain sequence to produce a gesture and to combine gestures into larger structures (i.e., words).

Signing and speaking involves many larger and smaller muscles and muscle groups put into play with extremely subtle differences in timing. Any skilled activity requires the appropriate sequencing of movements, and regulation of the degree of muscle contraction. These muscle actions occur with such rapidity that normal visual observation can hardly distinguish which of them are sequential, which simultaneous, which overlap in time.

A *coordinative structure* is a functionally defined unit of motor action: an ensemble of articulators that work cooperatively as a single task-specific unit across both abstract planning and concrete articulatory levels. Gestures are coordinative structures that involve an equivalence class of coordinated motions of several articulators to achieve a task [Browman and Goldstein, 1990: 300].

At the message or mental level, abstract units (segments) of language are discrete, static, and context-free. Consequently, language consists of sequences of discrete states of neuromuscular activity. However, the realization of segments is dynamic, context sensitive, and influences each other in coarticulation. *Coarticulation* is the extent to which individual phonetic elements are influenced by other elements before or after them so that the elementary form is altered slightly.

Different kinds of segment morphemes combine sequentially to form words: posture, preparatory, and activity. A posture describes how articulatory features (moveable parts) are configured. A preparatory movement achieves a basic posture from which activity movements will be performed to accomplish some task. Analogous to vowels

and consonants, activity morphemes are classified in motion and hold: motion is defined as a period of time during which some aspect of articulation is in transition, while a hold is a period of time during which all aspects of the articulation are in a steady state. In sign language, a preparatory segment is a morphological process known as *m-epenthesis* [Liddell and Johnson, 1989: 239].

The differences between visible movements are produced by muscle action. Such actions produce movement, but only certain combinations of the different movements constitute the words of human body movement. To describe accurately the phonology of human movement, phonological rules are required. The speech process involves more getting to segments than actually producing them. Hence, actions are more demanding than postures.

Morphology

An action is a functional unit, an equivalence class of coordinated movements that achieve some end. Actions can be achieved by a variety of means and entirely different body movements can achieve the same goal. The class of hand movements which carry an object contained in the hand towards the mouth could also be completed by leaving the hand static and moving the whole body and head so that the mouth moves closer to the object or with the entire body remaining stationary and the object moved towards the mouth by a second individual [Perrett at al., 1989: 109-110].

Phonemes selected and combined are put into morphemes (words and word classes) in another subsystem of language organization, morphology. Morphology provides the elements that syntax puts together into phrases and sentences. Words symbolize

classes of persons, things/objects, and actions/events. A complex movement combined with others forms a larger structure (i.e., coordinated patterns of gestures in time and space) that defines a word [Kelso at al., 1986: 31]. Words are articulatory programs composed of a few variable gestures [Studdert-Kennedy, 1987: 78]. In this sense, a human action corresponds to a sensory-motor word.

Syntax

The word syntax comes from the Greek syntaxis that means to arrange in order. Arranging things in order is the most fundamental requirement for the development of syntax. *Syntax* is a finite set of logical categories governed by complex interlocking rules capable of producing infinite combinations. Grammar constitutes a separate irreducible level of linguistic structure that is properly described without reference to meaning [Langacker, 1991: 515]. It has its own constructs, representations, and primitives. Getting from consciously produced signs to syntax is a matter of analysis (taking things apart) and planning (the ability to plan and assemble complex sequences of rapid motor actions), not synthesis (putting things together).

The Subject-Verb-Object (SVO) pattern of syntax is a reflection of the patterns of cause and effect: something doing something to something else. Syntax pairs relationships in the outside world with relationships within the brain. Syntax deals in networks not nodes, neural matrices not modules. Simultaneity (spatialization) is the primary generalization over spoken languages, since movement is constructed in three dimensions of space. Grammatical competence in movement involves spatial and sequential processes. Linguistic expressions are processed as if they were objects with internal structural configurations [Deane, 1991]. They are processed in terms of

certain basic image schemas (part-whole and linkage) critical to the recognition of the configurations that define complex physical objects. *Image schemas* are high-level embodied schemas that function as cognitive models of the body and its interaction with the environment. Image schemas are recurrent structures such as objects, shapes, figure-ground relations, source-path-goal, containment, compulsive force to move objects, and balance. Image schemata are representations of recurrent structured patterns that emerge from bodily experience. Cognition and language are built out of image schemata. Some basic image schemata include container, enablement, link, near-far, merging, center-periphery, compulsion, and part-whole [Johnson, 1987: 126].

Usually, there is a physical difference between a word and a sentence in spoken language, while they often look identical in sign language. Sentences represent very basic relationships. Arranging things in order is the most fundamental requirement for the development of syntax. Syntax emerges when sentences, relations between things and events, are made. According to the *spatialization of form hypothesis* [Lakoff, 1987: 283], grammar is ultimately spatial and the acquisition of grammatical competence occurs when linguistic information is routed to and processed by spatial centers in the brain. The spatialization of form hypothesis treats grammar as an image-schematic thought in which words, phrases, and sentences are endowed with an abstract structure grounded in immediate bodily experience of physical objects. Grammatical competence is critically represented in a brain region whose primary function is to represent the body schema and other high-level image schemata [Deane 1993: 278].

In most languages, the sequence of signals falls into a subject/predicate pattern. An action is represented by a word that has the structure of a sentence: the agent or subject is a set of active body parts; the action or verb is the motion of those parts. In many such words, the action is transitive and involves an object or another patient body part. The precise muscle timing (pre-syntax) makes it possible to produce countless actions that differ in great or small ways. A few prime symbols (S, NP, and VP) and a finite set of rules (S $\rightarrow$NP + VP and VP $\rightarrow$ V + NP) generate an infinite set of error-free symbol strings.

Movement Behavior

Observable movement is divided into involuntary and voluntary categories. Reflex (not voluntary) movements are elicited in response to some stimulus without conscious volition. Basic movements are inherent motor patterns which are based upon the reflex movements and which emerge without training. These movements form the basis for perceptual, physical abilities, and skilled movements.

The seven activities essential to the existence of primitive man are the basis upon which skilled movement is build. These actions, which are inherent in the human organism, include running, jumping, climbing, lifting, carrying, hanging, and throwing [Harrow, 1972].

Voluntary purposeful movement can be categorized as locomotor movement, non-locomotor movement (body in a stationary position), and manipulative movements. Locomotor movements change stationary state into ambulatory state by changing location (body moving in space from one point to another). Included in this subcategory are crawling, creeping, sliding, walking, running, jumping, hopping,

rolling, and climbing. Non-locomotor movements involve the limbs of the body or portions of the trunk in motion around an axis. Behaviors included in this subcategory are pushing, pulling, swaying, stooping, stretching, bending, and twisting. Manipulative movements are coordinated movements of the extremities usually combined with visual and tactile modality. This subcategory is concerned with movements of prehension and dexterity. Prehension is the combination of manipulative (flexion, gripping, inhibitory reflexes) and visual abilities with prehensive activity (reach for, grasp, and release grip). Dexterity implies a quick precise movement with hand and fingers (handling of blocks, cups, balls, and implements for drawing).

Study of movement behavior began in 1872 with the publication of Darwin's "The Expression of the Emotions in Man and Animals" [Darwin, 1872], a treatise on the origins and functions of facial and bodily expressions. Movement research ranges from emotional and psychophysiological dimensions (intrapsychic personality correlation to body motion) to interpersonal and cultural aspects of movement [Davis, 1973].

Naturalistic observation of infants and children indicates that movement patterns may be related to cognitive and personality development [Kestenberg at al., 1971]. There is evidence that one can diagnose schizophrenic symptoms from body movement patterns [Davis, 1970; Wolff, 1945]. Dyssynchrony of body parts in relation to one another is found in schizophrenic patients [Condon and Ogston, 1967]. Analyses of the psychological significance of various gestures or actions performed by a patient in psychotherapy suggests that movements can be immediate and visible reflections of

attitudes and feelings that are out of conscious awareness [Deutsch, 1952; Mahl, 1968]. While a simple correlation is made between anxiety and increased muscle tension, sophisticated analyses report more subtle possibilities such as between muscle activity and empathy, attention, and personality characteristics. Specific tensions in various parts of the body function as defenses against the experience and discharge of affect.

A given movement is interpersonally (reflects culture and role in a group) or intrapersonally significant (personality make-up or emotional state), where significance is that which the movement is associated with. The personality and cultural determinants of movement depend on which variables are predominant or characteristic for the individual or for the group. The parameters are aspects of movement that may be seen as the substrates of movement patterns in space. They include muscle tension patterns; expansion-contraction patterns (in breathing), weight placement; and body coordination variables (successive or simultaneous). For example, variables having to do with muscle tension and weight placement are related to emotion or personality dynamics. Intensity variables deal with movement qualities or variations in force, tempo, or rhythm. Space variables are direction, planes, or areas of space around the mover. Complex configurations of several movement parameters are gestalt aspects of body movement such as body attitudes, positions, and facial expression. The group variables refer to relationships between two or more people and include items such as group formation, orientation, or group synchrony.

The meaning of a given movement can only be determined by an analysis of the context: who does it, when the movement occurs, where, within what sequence of

interactions, and with what other behaviors in the communication stream [Birdwhistell, 1970]. As structural linguists analyze language, assume ignorance of the significance of movements and decipher the culture's meaning for any movement bit in relation to other same-size bits, then in relation to increasingly larger bits. There is rarely a relationship between the nature of the movement and its deciphered meaning, as with abstract mathematical symbols devised to represent operations. Usually, there is no resemblance between the character of the behavior and its meaning: kinesics is rarely onomatopoetic [Birdwhistell, 1970: 125].

A body can be bowed in grief, in humility, in laughter, or in readiness for aggression. However, each case shares an underlying common theme of containing oneself, whether it is associated with making oneself smaller in humility, holding oneself together in grief, preventing oneself from busting a gut laughing, or collecting oneself in preparation for attack. A humble bow may be done arms parallel to the side, head leading, the movement is smooth and controlled. Bowing in grief or in laughter may be done with arms tensely clutching one's sides, the whole body contracting, but with rhythm of breathing and shaking different in each case. Bowing in readiness to fight may involve a tight holding, but not the trembling or shaking of grief or laughter, and must of necessity involve outward focus. Knowledge of the situation would undoubtedly help in determining what exactly the particular bowing is associated with. However, the basic underlying pattern and the qualitative details of the movement itself may be the source of information as to its significance, as is an understanding of the context the movement occurs.

The developmental and functional significance of patterns are more a parsimonious explanation than some external process of association or learning of a cultural convention. For example, contracting the stomach muscles and holding one's breath can be seen as a way of coping with a painful experience, smiling is intrinsically related to pleasure in the sense that it is a free, widening, expanding movement, whereas frowning in distress or anger is a tight, contracting motion. Fundamental patterns within the organism are expanding in pleasure and contracting in displeasure. The intrinsic interpretation makes possible developmental and evolutionary explanations. However, the expression may take on other associations and evolve as a communicative signal for that with which it originally was associated. Vertical movements or stress on verticality within the individual mover has been associated with intrapsychic conflicts over control and self-assertion [Reich, 1949: 181].

Movement is intricately patterned at many levels and, with respect to movement, which occurs during speech, it may correlate with speech syntax, reflecting the beginning and ending of communication units and corresponding to various levels of speech structure.

Movement patterns relative to intrapsychic processes, emotion, or personality are intrinsic, whereas movement at the cultural level of face to face interaction and communication programs are arbitrary. However, the intrinsic relationship does not rule out the possibility that significance of movement is partly determined by the situation or by some extrinsic factors.

Movement is related to developmental processes, affect, intrapsychic and interpersonal dynamics, and cultural differences. It may be possible to observe

someone's movement over a limited period of time and, by noting a wide range of movement characteristics in varying degrees of detail, learn something about where the person is from (even the region of the country) [Birdwhistell, 1970: 208-210], the person's age, social status, sex, and class, as well as certain individual characteristics related to personality make-up. In addition, a subject's mental status [Davis, 1970; Wolff, 1945], mood [Clynes, 1970; Darwin, 1872], and even intelligence [North, 1971] may be analyzed. The practical application of movement analysis is in intelligence tests or psycho-diagnostic tests.

Another source of meaning to movements is body language, i.e., nonverbal communication. Speakers also control several paralinguistic systems such as facial expression and tone of voice to express and modify meaning.

Symbolic Representations

Symbolic representations of human activity are found in movement notation systems developed for dance and in linguistic studies about gesture and sign language. Dance notation systems are not accurate and designed for human reading and interpretation. There are many dance notation systems and among the most prominent are Labanotation [Hutchinson, 1977], Effort-Shape Analysis [Dell, 1971], and Eshkol-Wachmann [Eshkol, 1980].

Effort-Shape Analysis is the closest to a geometrical analysis of joint action and spatial patterns. Three types of movement are defined: a rotational movement in which a limb moves about its axis, a planar movement in which the longitudinal axis of the moving limb describes a plane, and a curved movement in which the

longitudinal axis of the moving limb describes a curved surface, usually a conic shape.

The symbols of notation systems may be seen as analogous to the notes and bars of music notation. Path and direction in space is comparable to pitch and tone in music, duration of the movement to duration of the note, and simultaneity of body parts moving in various directions to chords in music. However, there are special aspects of movement that have no clear counterparts in music, such as weight placement.

Evidence towards language embodiment grounded in spatio-motor system was found in linguistics. However, a symbolic representation has not been suggested. Linguists have proposed signed segments as movements and holds [Liddell, 1984], movements and locations [Sandler, 1986], movements and positions [Perlmutter, 1988]. Others have proposed that the common ground between signed and spoken languages will be found at the level of syllable [Wilbur, 1987], or that signed languages have no segments [Edmondson, 1987].

Reduce signing to phonetic writing systems [Stokoe, 1960; Stokoe at al., 1965] or to different two-dimensional representation is not satisfactory [Hockett, 1978]. Science of language and communication will be enabled by increasing sophistication in techniques of recording, analyzing and manipulating visible and auditory events electronically [Armstrong at al., 1995]. In this sense, this dissertation takes advantage of state-of-art motion capture systems to learn a linguistic structure for human activity.

Motor Primitives

Researchers in various disciplines have come close to the idea of primitives in human movement and primitives are the first step to a language. Indeed, recent work points to evidence that voluntary actions are made out of simpler elements, that are connected to each other either serially or in parallel (i.e., simultaneously) [Flash and Hochner, 2005; Hart and Giszter, 2004; Mussa-Ivaldi and Bizzi, 2000; Mussa-Ivaldi and Solla, 2004; Stein, 2005; Viviani, 1986]. This modularity provides the system with versatility and learning flexibility. To some scientists, motor primitives basically amount to motor schemas or control modules [Arbib, 1992; Jeannerod at al., 1995; Schaal at al., 2003], and they may be specific to a task. Their basic feature is that many different movements can be derived from a limited number of primitives through appropriate transformations, and that these movements can be combined through a well defined set of rules to form more complex actions (see for example the movemes of [Del Vecchio at al., 2003]). Primitives can be kinematic, dynamic, or kinemato-dynamic [Grinyagin at al., 2005; Hart and Giszter, 2004; Rohrer at al., 2002; Viviani, 1986], and may be extracted using statistical techniques like PCA (principal component analysis), HMM (hidden Markov models), and others.

At the behavioral level many have concentrated on reaching and grasping, gait and balance, posture and locomotion. Reaching movements appear to be coded in terms of direction and extent [Ghez at al., 1997], and appear to be composed of discrete sub-movements, all with a similar stereotypical, serially concatenated shape and overlapping in time [Fishbach at al., 2005; Pasalar at al., 2005; Roitman at al., 2004]. Motor primitives have also been examined for human and monkey grasping and

object manipulation. Prehension (such as lifting a full cup) consists of reaching, orienting the hand, and grasping. The three actions are executed as a unified coordinated complex act even though they can be combined in a variety of ways [Jeannerod, 1994]. In tasks such as grasping, not only must the positions of the fingers and motions be appropriately selected and preplanned but the forces exerted on the object must also be controlled to achieve the goal of the task while securing a stable grasp. Finger movements and forces have been decomposed into basic synergies based either on the idea of uncontrolled manifold or on inverse dynamics computations [Grinyagin at al., 2005; Kang at al., 2004]. Hand gestures also consist of primitives or more complicated sequences that can be decomposed into a series of more elementary units of activity [Jerde and Flanders, 2003]. Of particular interest are the motor synergies. These are simultaneous activations of several muscles that produce a torque about a joint or a force in a particular direction. EMG recordings from frog hind limb muscles have been analyzed to test whether natural behavior shows synergies among groups of muscle activities for an entire set of natural behaviors [Cheung at al., 2005; d'Avella at al., 2003; Hart and Giszter, 2004; Tresch at al., 1999]. Similar attempts have been made to find muscle synergies during human posture and locomotion [Ivanenko at al., 2005].

More recently using the technique of non-negative matrix factorization, muscle synergies during a postural task in the cat have been successfully identified [Ting and Macpherson, 2005]. Since several synergies were assumed to act on a given muscle, the total activation of that muscle is the sum of activations due to all the synergies. D'Avella and Bizzi [2005] employed a similar approach to extract amplitudes and

timing relationships among muscle activations during more natural behaviors in intact animals. A combination of synergies that were shared across behaviors and those that were for specific behaviors captured the invariance across the entire observed dataset. These results support a modular organization of the motor controller and that the motor output of these modules is combined to control a large set of behaviors.

Of special importance to our work is the finding that in the monkey cortex, electrical micro stimulation in primary motor and pre-motor cortex causes complex movements involving many joints and even several body parts [Graziano at al., 2002; Graziano at al., 2004]. These actions were very similar to gestures in the monkey's natural repertoire. Micro stimulation at each site caused the arm to move to a specific final posture. Thus there appears to be evidence for a cortical map of joint angles (or a cortical representation of limb or body postures). There appears also to be growing evidence that there is cortical coding not only of kinematic and dynamic variables but also of more global features (segment geometrical shape or the order of the segments within the sequence [Averbeck at al., 2003a; Averbeck at al., 2003b]).

We believe that the current debate on the exact nature of primitives is not very fruitful as far as artificial cognitive systems are concerned. The reason is that many different kinds of primitives are possible, and in our framework they will give rise to languages that are not identical. An important question in today's zeitgheist is to use large amounts of data to infer the global structure of the action space, or in our nomenclature, to learn the language with hyper-empiricism. In this sense, our approach is a holistic and empirical one that uses a very large number of

measurements of human actions. We capture these actions for many individuals with a motion capture system.

Motion Data Compression

The simplest way to reduce the size of motion data is by sampling the original data with frames equally spaced. Naka at al. [1999] present a method for the compression and decompression of motion streams. The compression method performs uniform sampling in the time axis and further quantization in the floating-point byte representation of values.

Togawa and Okuda [2005] perform key frame detection in the positions of the joints in 3D space (i.e., translational data). The less important frames are decimated iteratively according to a cost function. This function is computed as the sum of all joint position distances between consecutive frames. The frame with the lowest cost is the less important and, consequently, decimated.

Non-uniform (adaptive) sampling involves the identification of points irregularly spaced in time. Chenevière and Boukir [2004] propose a non-uniform segmentation approach using a deformable model and active contour fitting. The active contour is a subset of the points in the motion trajectory. Initially, the contour has only the extremities of the motion trajectory. New vertices are inserted iteratively through an optimization step which minimizes energy cost in the contour segment with highest approximation error.

Curve fitting approaches for motion data compression find representative points that characterize the motion trajectory. *Polygonal approximation* [Etou at al., 2004; Latecki and Lakämper, 1999] is a method of curve fitting where a curve is

represented by a piecewise-linear polygonal line. Lim and Thalmann [2001] use a batch curve simplification algorithm to identify key postures in human motion data modeled as high-dimensional curves of rotational data. A simplification algorithm generates an approximation of a curve as a smaller number of line segments. In our compression experiments, we use an online version of this curve simplification algorithm applied to two-dimensional curves representing each DOF in the motion data. Another curve fitting technique is *polynomial interpolation* [Saux, 1999; Sudarsky and House, 1998], which approximates the low frequencies of an input signal using spline or B-spline curves.

Learning through Imitation

Another approach towards motion data compression and segmentation is dimensionality reduction. Assa at al. [2005] embed the high-dimensional motion curve in a low-dimensional Euclidean space. The dimensionality reduction is performed in affinity matrices by a non-linear optimization process: replicated multi-dimensional scaling [McGee, 1978]. A number of affinity matrices are defined to describe different aspects of inter-pose distance or similarity. The local extreme points which are not close to each other are identified as key poses of the motion. Barbič at al. [2004] present methods based on statistical properties of the motion. They consider the intrinsic dimensionality from PCA of a local model of the motion and the local change in the distribution of poses. Sidenbladh at al. [2002] construct a low dimensional linear model of the human motion. They use PCA to reduce the dimensionality of the time series of joint angles. The movement data is structured into a binary tree using the coefficients with larger variance in higher levels of the tree.

Jenkins and Matarić [2003] use dimensionality reduction to extract motion primitives (spatio-temporal structure) with an extension of the Isomap algorithm. The algorithm performs eigenvalue decomposition on a similarity matrix computed as a geodesic distance.

Fod at al. [2002] present a method for automatically generating a set of movement primitives from human multi-joint movement. They segment motion according to angular velocity at points where more than one DOF has a zero velocity crossing. Primitives are found by k-means clustering the projection of high-dimensional segment vectors onto a reduced subspace. While they are more interested in the definition of basic movements which serve to compose movement through linear combination, we focus on atomic representations. They introduce two kinetological principles, consistency and completeness, referred in this dissertation as reproducibility and reconstructivity, respectively. However, the evaluation of action representations according to them and the other principles introduced here are original results of our work.

Reduced-space approaches [Kovar and Gleicher, 2004] infer a parametric space that, in general, lacks a correspondence to intrinsic properties of the movement. In Figure 2.1, we show several performances of a kick action towards nine different target locations regularly spaced in a two-dimensional vertical plane. Each point corresponds to one instance of a kick action and its color represents the target location. The kick motions are reduced to a two-dimensional space (responsible for more than 95% of the variation). The 2D projections of kick action instances are presented in the graph as colored points. We also show a convex hull for instances of

each of the nine target locations. Note that the parametric subspace (i.e., convex hull) associated with some target locations overlap. This is counter-intuitive since kick motions for different target locations correspond possibly to the same point in the reduced space.

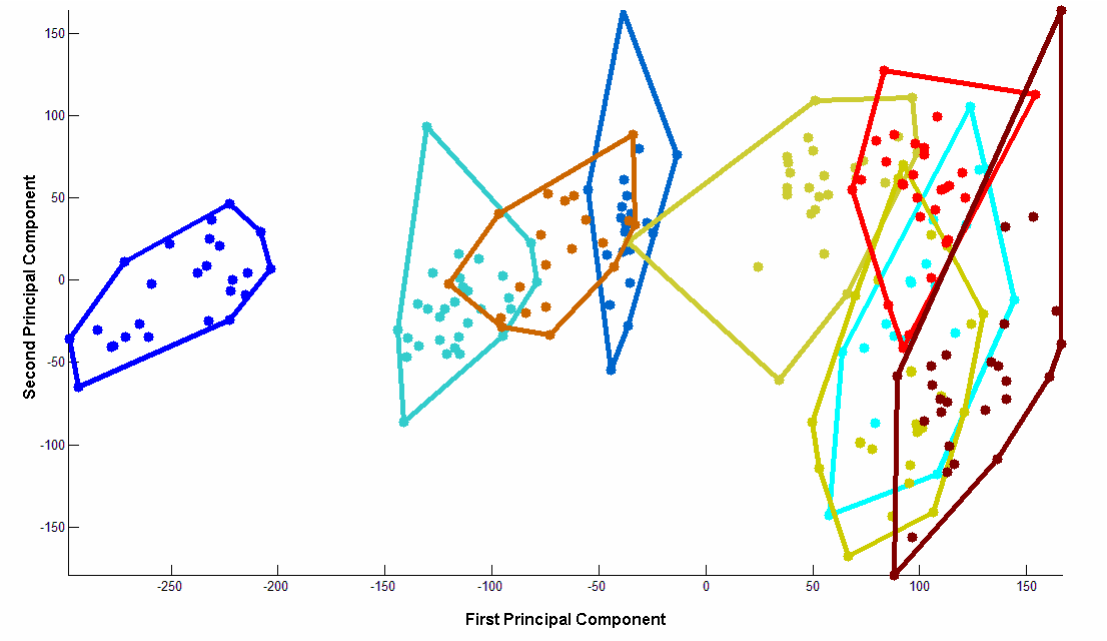


Figure 2.1. Reduced dimensionality space representation.

The previous techniques lack the intuitive interpretation for the extracted segments. Intuition appears in segmentation when characteristic features are used. In general, these features are spatio-temporal features, such as the curvature of 3D trajectories [Caelli at al., 2001; Rao and Shah, 2001] and kinematic features.

Mori and Uehara [2001] use kinematic motion primitives to discover association rules representing dependency between body parts during movement. Their segmentation is based on the velocity of joint points in the Cartesian space, where each axis is treated independently and integrated afterwards.

Ilg at al. [2004] extracts movement primitives based on key events defined by zeros of the velocities in selected degrees of freedom. The movement primitives are characterized by the angular displacement between key events. Robust identification of primitives is accomplished by a dynamic programming algorithm. The algorithm aligns the features obtained from a search window to a prototypical movement primitive learned previously.

Kahol at al. [2004] introduce a methodology for automatically parse discrete gestures from a continuous stream of modern dance motion (gesture segmentation). They use inertial factors derived from velocity, acceleration, and mass. These factors are integrated according to the hierarchy of the human body. The local minimum in total body force is used to detect segment boundaries. In their approach, all joints are treated as a single feature and the gesture movements segmented are not atomic.

Nakazawa at al. [2002] use human dance motions and consider local minimum velocity only of the end effectors to represent whole body motion. They measure similarities of motion segments according to a dynamic programming distance of the trajectories in 3D space and cluster these with a nearest-neighbor algorithm. Kuniyoshi at al. [1994] generate complex activities from observation of human action based on abstraction and symbolization. Samejima at al. [2002] present another approach on learning from demonstration. Inamura at al. [2002] introduces the mimesis model where motion patterns are analyzed into motion elements and associated with proto-symbols. HMMs are used to recognize motion behaviors as symbols. For motion generation, genetic algorithm is used with the HMM likelihood as a fitness function.

Semantic Gap

The sensory-motor projection of primitive words leads to language grounding. Language grounding for verbs has been addressed by Siskind [2001] and Bailey at al. [1998] from the sole perspective of perception and action, respectively.

There are several approaches towards bridging the semantic gap between low-level features and high-level concepts. Relevance feedback [Rui at al., 1998] is an interactive approach for content-based image retrieval. The relevant images are selected according to user feedback and low-level features extracted from each image in a database. Hidden annotation [Cox at al., 2000] further extends these features by including manually Boolean semantic attributes (e.g., person, city, animal) in the relevance inference. Usually, image databases are only partially annotated due to the heavy manual labor involved. Active learning [Zhang and Chen, 2002] aims to determine which subset of the database should be annotated. In this sense, our approach is a step towards fully automatic annotation. Given the motion information, each action is automatically converted into our symbolic linguistic representation and linked to the corresponding concept for further processing. Usual text search engines and other symbolic manipulation techniques could be used for the retrieval of multimedia information.

Automatic Computer Animation

In this section, we review representative work related to automatic motion synthesis problems. The motion interpolation problem was addressed by Wiley and Hahn [1997] with a linear interpolation method to create parameterizations of human

activities. Radial basis functions were used by Rose et al. [1998] to interpolate between different styles of the same action.

In motion splicing, a naïve DOF replacement consists in just swapping data between two motions [Ikemoto and Forsyth, 2004; Perlin, 1995]. Rose et al. [1996] generated motions splicing the DOFs of the right arm (e.g., wave and salute) and walking. However, spatio-temporal correlations are ignored and, consequently, unrealistic motions were produced. Ashraf and Wong [2000] addressed the correlations by using frame space interpolation. They show how upper-body action and lower-body locomotion can be treated independently during blending. Ko and Badler [1996] used inverse dynamics to adjust walking motions so a character would stay in balance and reasonable joint torques. Chai and Hodgins [2005] reconstruct a whole-body motion from a subset of joint positions. Heck et al. [2006] generate motions for combinations of *n* locomotion and *m* upper-body action. They propose a method for splicing upper-body action and lower-body locomotion. Their method identifies and enforces temporal and spatial relationships between upper body and lower body.

The transitioning problem is usually solved by graph-based approaches [Arikan and Forsythe, 2002; Lee et al., 2002]. A motion graph [Kovar et al. 2002] is a directed graph that represents a collection of motion sequences, where each node is a motion frame and each edge connects similar frames. Motion synthesis is achieved by computing paths in the motion graph. A move tree [Menache, 2000] is a graph where edges are segments of motions and nodes are transitions between motions. While motion graphs are created automatically, transitions of move trees are chosen manually.

Markerless Motion Capture

The mapping from low-level visual features to human movement can be achieved implicitly or explicitly through motion capture. Motion capture is the process of recording real life movement of a subject in some digital geometric representation (e.g., Cartesian coordinates or Euler angles). Optical motion capture uses cameras to reconstruct the body posture of the human performer. One approach employs a set of multiple synchronized cameras to extract markers placed in strategic locations on the body, as shown in Figure 2.2.

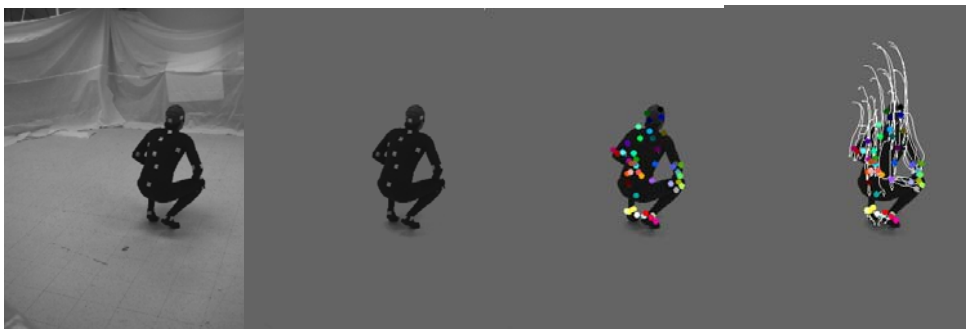


Figure 2.2. Optical motion capture.

A more flexible method, markerless monocular (single camera) motion capture (MMMC), avoids the use of markers and extends the capabilities of such systems to any input video. A model-based approach [Chen at al., 2005] for MMMC uses a 3D articulated model of the human body to estimate the posture such that the projection of the model fits the image of the performer for each frame. Data driven techniques [Lee at al., 2002] use a motion database to help in the reconstruction of the motion in the video. The motion database is pre-processed to create connecting transitions between similar poses according to kinematic features.

Given a video featuring human actions, a MMMC system extracts the human movement from visual features such as silhouettes. Joint angles are computed for all DOFs in a hierarchical body model. These joint angle functions are the initial input for segmentation in our linguistic framework.

Action Recognition

Stuart and Bradley [1998] find interpolation sequences between pairs of body postures using A* search in a set of transition graphs built from corpora of human movement. These graphs capture the progressions of a single joint in the corpus.

HMMs are vastly used to characterize movement sequences [Yang at al., 1997]. Alon at al. [2003] estimate a finite mixture of HMMs using an expectation maximization formulation. In this approach, segments are partially assigned to all clusters corresponding to HMMs. Brand and Hertzmann [2000] extend HMM with a multidimensional style variable used to vary its parameters. They learn motion patterns from a set of motion sequences. HMMs are essentially probabilistic finite state automata. In this sense, a stochastic context-free grammar (SCFG) is a generalized model, which relaxes some structural limitations. Ivanov and Bobick [2000] use a single SCFG to parse activities and interactions between multiple agents. Wang at al. [2001] present a gesture segmentation approach for human gestures represented as 2D trajectories of projected hands. The segmentation involves finding the local minima of velocity and local maxima of change in direction. The segments are hierarchically clustered into classes associated with symbols using HMM to compute a metric. A small lexicon is inferred from the symbolic sequence through a language acquisition approach. The lexicon is induced for a single movement

stream/string and, consequently, involves only sequential learning which suffers from the overgeneralization problem.

Mörchen at al. [2005] present a framework to discover movement patterns from EMG and kinematic measurements represented as multivariate time series. The kinematic time series are reduced to primitive patterns by manual clustering with Emergent Self-Organizing Maps and no time information. The same consecutive primitives are merged into intervals corresponding to symbolic states. They assume all actuators are participating equally in the action. While they consider all aspects of movement at the same time (total body movement) to find coincident intervals, our approach identifies the relevant actuators involved in the movement automatically and considers actuators independently. Further, in their approach, the pattern events discovered are sparse and cannot be used for the reconstruction of the movement.

To the best of our knowledge, no approach modeling human motion learns the set of actuators involved in an action. Usually, they consider a fixed set of actuators and, since our method induces the appropriate actuator set for each action, a comparison between our technique and others is unfeasible.

Grammatical Inference

Here we pose the morphology of human activity as a grammatical inference problem. *Grammatical inference* concerns the induction of the grammar of a language from a set of labeled sentences. The grammar inference consists in learning a set of rules for generating the valid strings that belong to the language. The target grammar usually belongs to the Chomsky hierarchy of formal grammars. There exist several methods

for learning regular grammars, context free grammars (CFGs), and stochastic variations [Parekh and Honavar, 2000].

Regular grammars and context free grammars cannot be induced only from positive examples [Gold, 1967]. However, several heuristic techniques learn approximations to the target grammar. The SNPR algorithm [Wolff, 1988] learns syntagmatic elements (sequences) and paradigmatic elements (sets) from minimal elements which are perceptual primitives (e.g., letters or phonemes). Each element corresponds to a rule in the learned grammar. The learning involves the concatenation of the most frequent pair of contiguous elements.

Sequitur [Nevill-Manning and Witten, 1997] is an algorithm that infers a hierarchical structure from a sequence of discrete symbols. Sequitur infers a grammar, where each repeated subsequence gives rise to a rule and is replaced by a non-terminal symbol. The algorithm constrains the grammar with two properties: digram uniqueness and rule utility. The algorithm operates by enforcing these constraints on an online stream.

Current approaches [Nevill-Manning and Witten, 1997; Solan at al., 2005; Wolff, 1988] account only for sequential learning and not for the parallel learning, inspired by associative learning, we introduce. We define sequential learning as the technique able to infer the structure of a single sequence of symbols A. The learning algorithm induces a CFG corresponding to the structure of the string representing the movement. This structure corresponds to a forest of binary trees, as shown in Figure 2.3, where each node in a tree is associated with a context-free grammar rule in a normal form. Initially, the sequential learning algorithm computes the number of

occurrences for each different digram in the string A. A *digram* is a pair of adjacent symbols. A new grammar rule $N_c \rightarrow \alpha\beta$ is created for the digram $\alpha\beta$ with the current maximum frequency. The algorithm replaces each occurrence of $\alpha\beta$ in the string A with the created non-terminal N_c. The whole procedure is repeated until digrams occur more than once. As an example, the set of rules inferred for the CFG displayed in Figure 2.3 is $\{N_1 \rightarrow \mathbf{AB}, N_2 \rightarrow \mathbf{CD}, N_3 \rightarrow \mathbf{EF}, N_4 \rightarrow \mathbf{B}N_1, N_5 \rightarrow N_2N_3, N_6 \rightarrow N_5\mathbf{G}, N_7 \rightarrow N_6N_4\}$. A sequential learning algorithm keeps merging adjacent root nodes into single rules and, consequently, overgeneralization happens when "unrelated" rules are generalized.

Figure 2.3. A CFG shown as a binary tree forest.

Grammar Systems

Variants of the classical models in formal language theory are used to specify non-determinism in computing devices with notions such as distribution, parallelism, concurrency, and communication. A grammar system consists of several grammars (components) that work together generating a common symbolic state represented by a finite set of strings. The components of the system change the state through rewriting and communication.

We use grammar systems as a formal model to learn the morphological structure of human actions. Other formalizations in natural language processing such as synchronous grammars are avoided. The reason is that these grammars are used in

machine translation to correspond structures in different languages with the same meaning, while in human motion modeling, different actuators play different roles executing synchronously distinct unrelated motor programs.

The most important models of grammar systems are cooperating and parallel grammars. Cooperating Distributed Grammar Systems (CDGS) have components working sequentially [Csuhaj-Varjú and Dassow, 1990; Meersman and Rozenberg, 1978]. Only one component is active at any moment. Therefore, the components take turns in rewriting a common sentential form according to a certain cooperation protocol. Colonies are a simplification of CDGS where the components are regular grammars generating finite languages. Sosík and Štýbnar [1997] train a Neural Pushdown Deterministic Automaton (NPDA) with sequential access to a set of positive and negative sequences in some language. The NPDA model requires preliminary information about the expected size of the inferred grammar, since the topology of the NPDA does not change during the training. They extract a colony from the trained NPDA with a heuristic algorithm after a hierarchical clustering in the space of neuron states.

A Parallel Communicating Grammar System (PCGS) consists of several grammar components working simultaneously in synchronization [Păun and Sântean, 1989]. The component grammars rewrite their own sentential forms in parallel. They communicate by exchanging their current sentential forms among each other. The requested string becomes part of the sentential form of the receiving grammar. In a returning mode, after sending their partial solutions to others, the components are reset to their axioms and start a new computation. The language generated by the

system is the language generated by a distinguished component of the system (master grammar) with the help of the others.

The assumption that communication takes a single step and components continue computation without waiting for the end of communication is not reasonable. Fernau [2001] discusses a variant of PCGS with terminal transmission and right-linear components. In this model, the communication is constrained only to the transmission of terminal strings. Therefore, queried components have only terminal strings as sentential forms by definition. An inference algorithm for this model is proposed which uses additional structural information about communication (sentences with query symbols) and the component languages are learned separately with special care for the master component.

Chapter 3: Kinetology

Human movement is a biological phenomenon which consists in the voluntary motion of the human body. The understanding of the biomechanical bases and description of human movement contributes to the improvement of this capacity in humans. On the other hand, these motor aspects of human movement have important applications to artificial systems in the synthesis and analysis of movement.

Motion synthesis is the generation of movement for animation characters with a realistic appearance which aims to avoid unnatural and mechanical artifacts. Mostly, realistic motion synthesis is based on real examples coming from motion capture. Human motion capture usually corresponds to a very large amount of data. Segmentation, the extraction of key postures or motion primitives, summarizes the motion content and results in the compression of motion data.

The precise exemplar movements are constrained only to the ones stored in a motion database library. Novel realistic movement either needs to be captured or adapted from previously recorded motion. Adaptation involves the reuse of motion segments, manipulation of motion attributes, and sequencing (concatenation) of movement according to physics laws. This way, any representation should assist in those tasks and be able to reconstruct the original and adapted movement.

On the other hand, motion analysis relates to perception and involves the parsing of visual information into action representations ranging from optical flow to stick figure models. These representations are used to uniquely identify the action performed in a video. Therefore, an action representation should be able to select among different activities and to reproduce the same structure for different

performances of the same action. Furthermore, an action representation should be based on primitives robust to variations of the image formation process. In this sense, camera view-invariance is a desired property for representations dealing with motion analysis.

Adequate primitives and segmentation must consider both generation and perception of movement. One reason is that motion synthesis involves the generation of animation satisfying a realistic criteria based ultimately on perception. On the other hand, motion analysis should map the parsed structure from video into a representation which should regenerate the original observed motion. Furthermore, an integrated approach would allow imitation, an important component of an artificial cognitive system [Matarić, 2002]. Therefore, the research problems of motion synthesis and motion analysis should be combined and based on common representations.

Action units in behavior are all organized within a clearly definable narrow time window or temporal segment. This temporal segmentation appears to represent a basic property of the neuronal mechanisms underlying the integration and organization of successive events [Kien, 1992: 19]. If words are formed from simultaneous combinations of gestures, the perception somehow finds these elements in the movement signal. The signal cannot be divided into a neat sequence of units and the patterns associated with a particular segment vary with the phonetic context. The lack of invariant segments in the signal matching the invariant segments of perception constitutes the *anisomorphism paradox* [Studdert-Kennedy, 1985: 142].

An initial step in our linguistic framework is to find basic primitives for human movement. These motion primitives are analogous to phonemes in spoken language. While phonemes are units of phonic origin (sounds), the motion primitives are units of kinetic origin (movement) that we refer as *kinetemes*. These atomic units are the building blocks of a foundational system for human movement denoted as *kinetological system*. The problem addressed in this chapter concerns the representation of human movement in terms of atomic sensory-motor primitives.

In this sense, *kinetology* is dedicated to the study of systems of movement as the foundations for a kinetic language. In addition to a geometric representation for 3D human movement, a kinetological system consists of segmentation, symbolization, and principles. We introduce a kinetological system with five principles on which such a system should be based: compactness, view-invariance, reproducibility, selectivity, and reconstructivity.

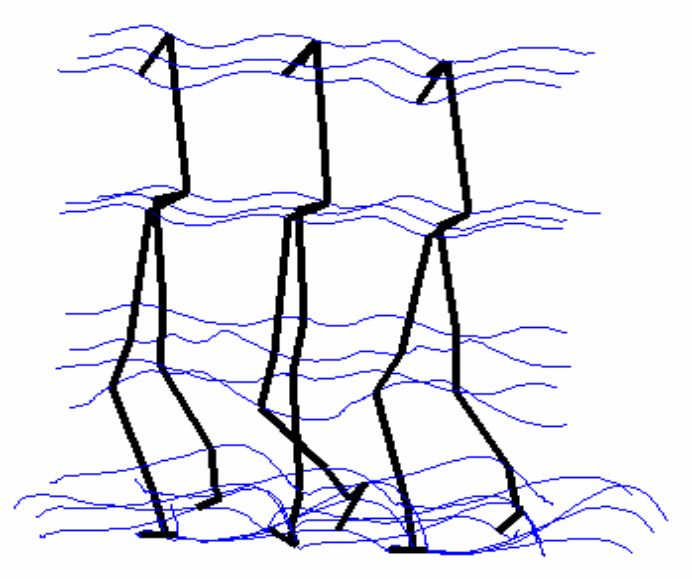
We propose sensory-motor primitives and demonstrate their kinetological properties. Further evaluation is accomplished with experiments on compression and decompression of motion data. To represent human movement satisfying the above requirements, we consider whole body movement associated with general human actions. Although we consider whole body movement, each DOF is treated independently. An initial 3D geometric representation for human movement is assumed as input towards the computation of our sensory-motor representation. Actual movement data is analyzed in the process of evaluating the proposed kinetological system according to its principles.

Geometric Representation

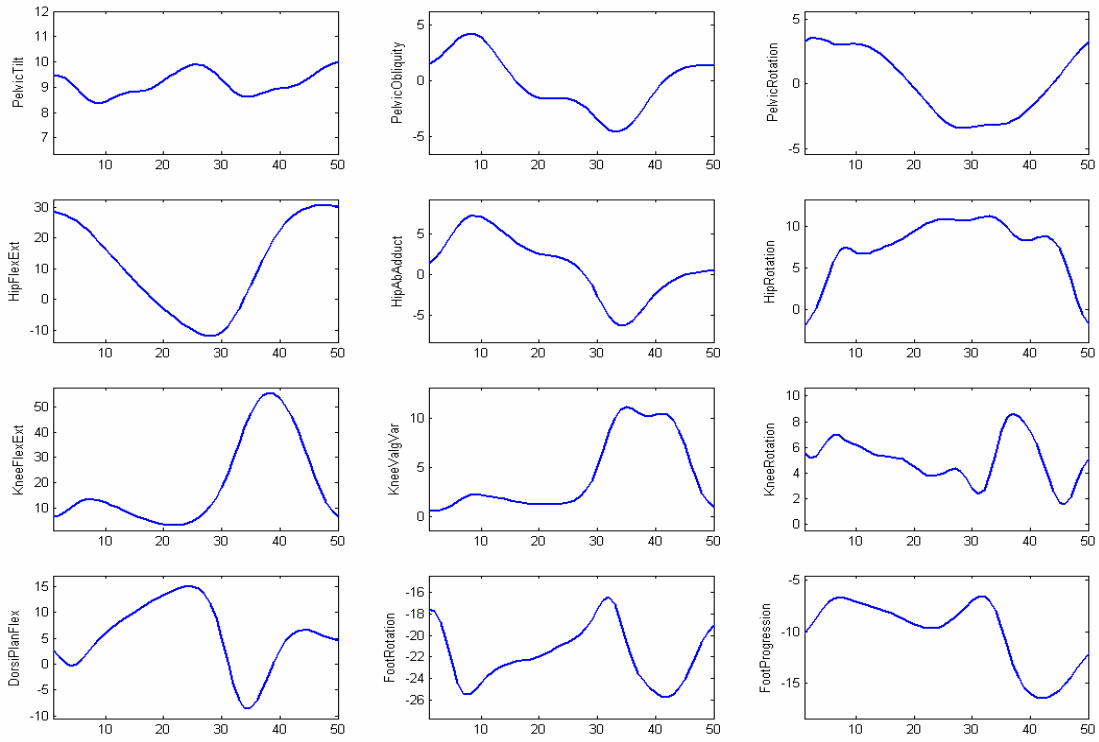
A model for the human body which considers only rigid articulated movement consists of a skeleton. A *skeleton* is defined as a set of rigid body parts connected through joints. Formally, the topology of a skeleton is modeled as a graph where vertices correspond to body parts and edges are associated with joints. A posture is the geometric configuration of the skeleton at one instant. Human movement consists in the continuous time variation of postures. There are two basic 3D geometric representations for whole body movement: external and internal.

The *external representation* consists of a set P of points in the human body, as shown in Figure 3.1a. At an instant t, a point $p_i \in P$ is associated with the corresponding 3D Cartesian coordinate $[X_i(t), Y_i(t), Z_i(t)]$. The whole human movement is fully determined if at least three points in each rigid body part are included in the representation. This way, a local coordinated system can be defined for each body part.

The degrees of freedom for a joint can be recovered from the transformation relating the two local coordinated systems corresponding to the adjacent body parts. The *internal representation* of human movement may use Euler angles to specify the rotational degrees of freedom of each joint. The internal system describes human movement with a set Q of joints, where a joint $q_j \in Q$ is associated with Euler angles $\phi_j(t)$, $\theta_j(t)$, and $\psi_j(t)$ at instant t, as shown in Figure 3.1b.



(a) External representation.



(b) Internal representation.

Figure 3.1. Three-dimensional representations of human movement.

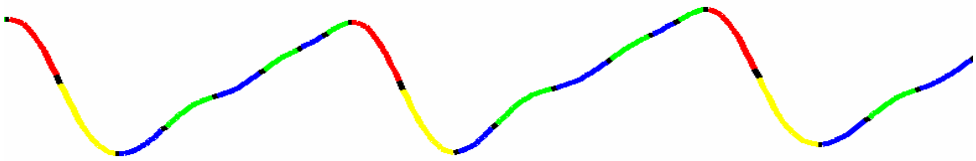
The internal representation makes explicit use of embodiment through the topological specification of a skeleton. The topological graph of a skeleton is defined as a tree where the root resembles the human vestibular system. This system provides measurements about global movement and orientation in space for humans. The internal representation is analogous to the *proprioceptive system* which monitors movement and is responsible for *kinesthesia*: the sense of body position awareness.

Segmentation

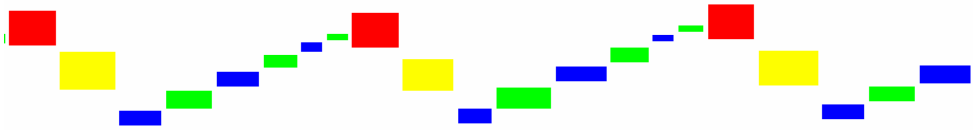
The input for our kinetological system is real human motion obtained with a motion capture system. Each DOF i in a model for the articulated human body, refered as *actuator*, corresponds to a time-varying function J_i. The value $J_i(t)$ represents the joint angle of a specific actuator i at a particular instant t. In kinetology, our goal is to identify the motor primitives (segmentation) and to associate them with symbols (symbolization). This way, kinetology provides a non-arbitrary grounded symbolic representation for human movement. While motion synthesis is performed by translating symbols into motion signal, motion analysis uses this symbolic representation to transform the original signal into a string of symbols used in the next steps of our linguistic framework.

Automatic segmentation is the decomposition of action sequences into movement primitives. Theses primitives are atomic elements with characteristic properties that stay constant within a segment. This concept of motion primitives differ from the one associated with behavioral basis [Matarić, 2002], which are used for composition of movement through linear combination. To segment human movement, we consider each actuator independently. An actuator is a 3D point or a joint angle describing the

motion in an external or internal representation, respectively. Each joint angle is represented as a one-dimensional function over time. We associate an actuator with a joint angle specifying the actuator's original 3D motion according to an internal geometric representation as shown in Figure 3.2a. The segmentation process assigns one state to each instant of the movement for the actuator in consideration. Contiguous instants assigned to the same state belong to the same segment, as Figure 3.2b shows.



(a) Geometric representation.



(b) Segmentation.

(c) Symbolization.

Figure 3.2. Kinetological system.

We define a state according to the sign of derivatives of a joint angle function. In our segmentation method, we use angular velocity J' (first derivative) and angular acceleration J'' (second derivative), as shown in Figure 3.3. This leads to a four-state system: positive velocity/positive acceleration ($J'_i(t) \geq 0$ and $J''_i(t) \geq 0$), positive velocity/negative acceleration ($J'_i(t) \geq 0$ and $J''_i(t) < 0$), negative velocity/positive acceleration ($J'_i(t) < 0$ and $J''_i(t) \geq 0$), and negative velocity/negative acceleration ($J'_i(t) < 0$ and $J''_i(t) < 0$). It is worth noting that a kinetological system can be defined

in both complex (considering higher order derivatives such as jerk) and simple ways. A simpler system could have used only the first derivative. In that case, we would have only two states: positive velocity ($J'_i(t) \geq 0$) and negative velocity ($J'_i(t) < 0$). Higher order derivatives increase the amount of segmentation, adding complexity to the description of the movement. The number 2^h of possible states depends on the order h of the highest derivative used.

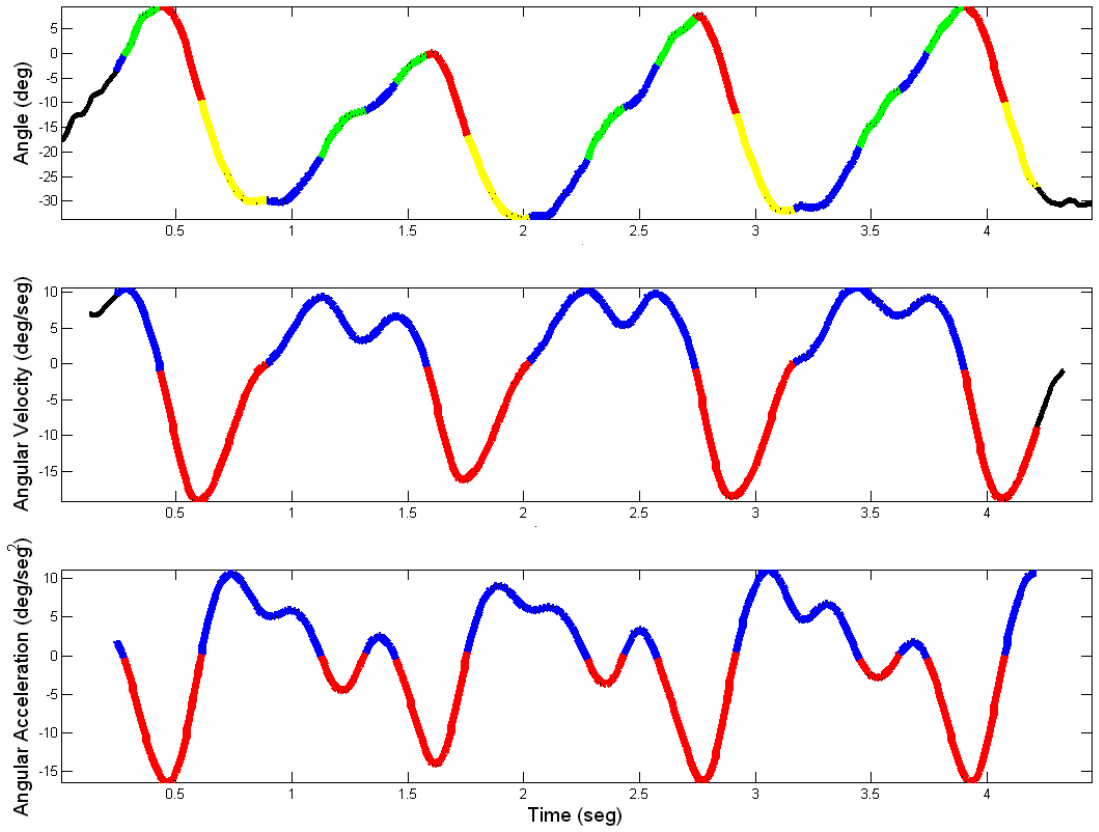


Figure 3.3. Angular derivatives used in our segmentation method.

The representation has a qualitative aspect, the state of each segment, and a quantitative aspect corresponding to the time length and angular displacement (i.e., the absolute difference between initial joint angle and final joint angle) of each segment. Once the segments are identified, we keep these three attribute values for

each segment: the state, the time length, and the angular displacement. Each segment is graphically displayed as a filled rectangle, where the color represents its state, the vertical width corresponds to angular displacement, and the horizontal length denotes the time length, as Figure 3.2b shows. The four colors used to depict a four-state kinetological system are blue for positive velocity/positive acceleration segments, green for positive velocity/negative acceleration segments, yellow for negative velocity/positive acceleration segments, and red for negative velocity/negative acceleration segments. In a two-state kinetological system, the two colors used are blue for positive velocity segments and red for negative velocity segments. Given a compact representation, the attributes are used in the reconstruction of an approximation for the original motion signal and in the symbolization process.

Symbolization

The kinetological segmentation process results into atoms observing some natural variability. Our goal is to identify the same kineteme amidst this variability. The symbolization process consists in associating each segment with a symbol such that segments with the same state corresponding to different performances of the same motion are associated with the same symbol. Symbolization amounts to classifying motion segments such that each class contains variations of the same motion. This way, each segment is associated with a symbol representing the cluster that contains motion primitives with a similar spatiotemporal structure, as Figure 3.2c shows. Hierarchical clustering, using an appropriate similarity distance for segments with the same atomic state, offers a simple way to perform symbolization.

Another way to perform symbolization is to compute a graph, where the set of vertices corresponds to all segments with the same atomic state. There exists an edge between two vertices in the graph if the similarity distance between the two corresponding segments is less than a threshold value. The similarity distance is the absolute difference between the time normalized versions of the joint angle functions associated with the segments. The symbolization clusters are the connected components of the similarity graph.

A probabilistic method to achieve symbolization is model-based probabilistic clustering. Different from model-based clustering, we also used a generalized probabilistic clustering algorithm to classify segments for each joint angle independently. A segment is represented as tuple (α, d, t), where α denotes the atomic state, d corresponds to the angular displacement, and t is the time length. The movement corresponding to a specific joint angle is segmented into a sequence of m atoms (α_j, d_j, t_j) for j = 1, …, m. Our algorithm partitions the 2D parametric space concerning the quantitative attributes (d, t) into regions of any shape.

Initially, we compute probability distributions over the 2D parametric space, as shown in Figure 3.4. We find one distribution P_α for each of the possible states by considering only the atoms where $\alpha_j = \alpha$. Each atom (α_j, d_j, t_j) contributes with the probability modeled as a Gaussian filter $h(k_1, k_2)$ centered at (d_j, t_j) with size $(2W_D + 1)\times(2W_T + 1)$ and standard deviation σ. This way, the probability distribution is defined as

$$P_\alpha(d,t) = \frac{1}{m} \sum_{\substack{j=1 \\ \alpha_j=\alpha}}^{m} \sum_{\substack{k_1=-W_D \\ d_j+k_1=d}}^{W_D} \sum_{\substack{k_2=-W_T \\ t_j+k_2=t}}^{W_T} h(k_1,k_2).$$

Once the probability distribution P_α is computed, each local maximum is associated with a class. This way, the number of clusters is selected automatically. The partitioning of the parametric space is performed by selecting a connected region for each cluster c associated with a local maximum p_c. For a cluster c, we find the minimum value v_c such that the region r_c in the parametric space satisfying $P_\alpha(d, t) > v_c$ contains only the peak p_c and no other.

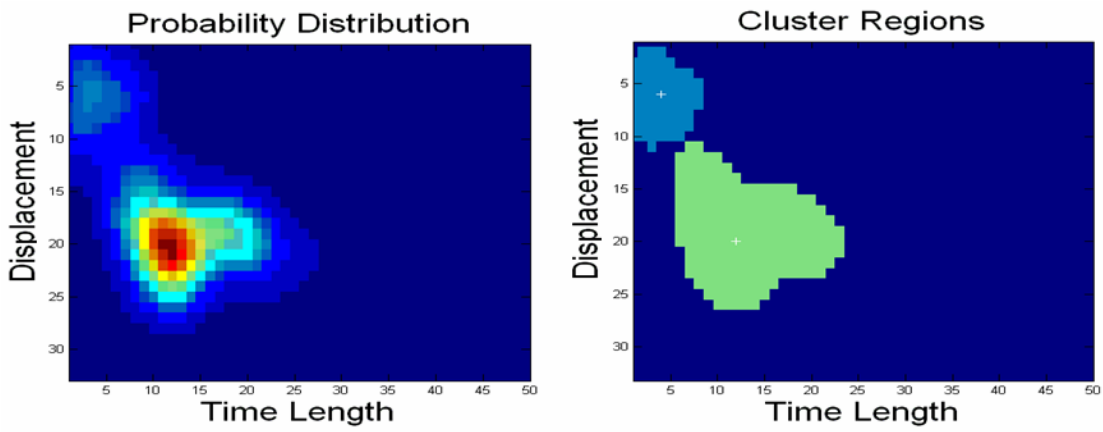


Figure 3.4. A generalized probabilistic clustering method for symbolization.

Each sample atom (α_j, d_j, t_j) is assigned to the cluster c which maximizes the expected probability

$$e_c(d_j, t_j) = \sum_{k_1=-W_D}^{W_D} \sum_{k_2=-W_T}^{W_T} h(k_1, k_2) \bullet R_c(d_j + k_1, t_j + k_2),$$

where R_c is a binary matrix specifying the connected region r_c corresponding to the cluster c. This probabilistic clustering algorithm uses a more general model than standard probabilistic clustering techniques.

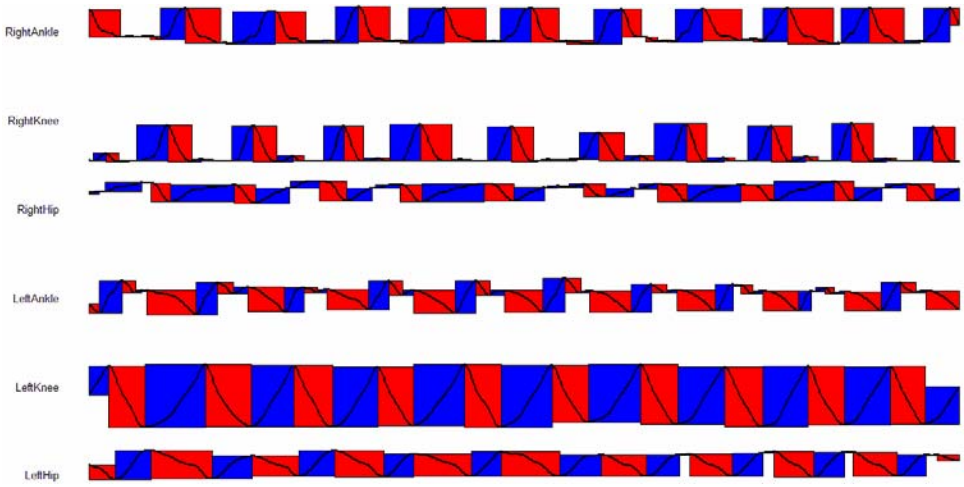


Figure 3.5. Segmentation of human motion.

Given the segmentation for a motion data, as shown in Figure 3.5, the symbolization output is a string of symbols for each actuator in the body. This set of strings for the whole body defines a single structure, denoted as actiongram, shown in Figure 3.6. An actiongram A has n strings $A_1, \ldots, A_n$. Each string A_i corresponds to an actuator of the human body model and contains a possibly different number of m_i symbols. Each symbol $A_i(j)$ is associated with a segment and its attributes.

Principles

Besides sensory-motor primitives, we suggest five kinetological properties to evaluate our approach and any other: compactness, view-invariance, reproducibility, selectivity, and reconstructivity. We describe these principles in detail and demonstrate that our segmentation method and primitives possess these properties.

Compactness

The compactness principle relates to describing a human activity with the least possible number of atoms to decrease complexity, improve efficiency, and allow

compression. We achieve compactness through segmentation, which reduces the representation's number of parameters. We implemented our segmentation approach as a compression method for motion data, tested our compression efficiency algorithm on several different actions, and recorded a median compression rate of 3.698 percent of the original file size for all motion files. We achieved the best compression for actions with smooth movement. Further compression could be achieved through symbolization.

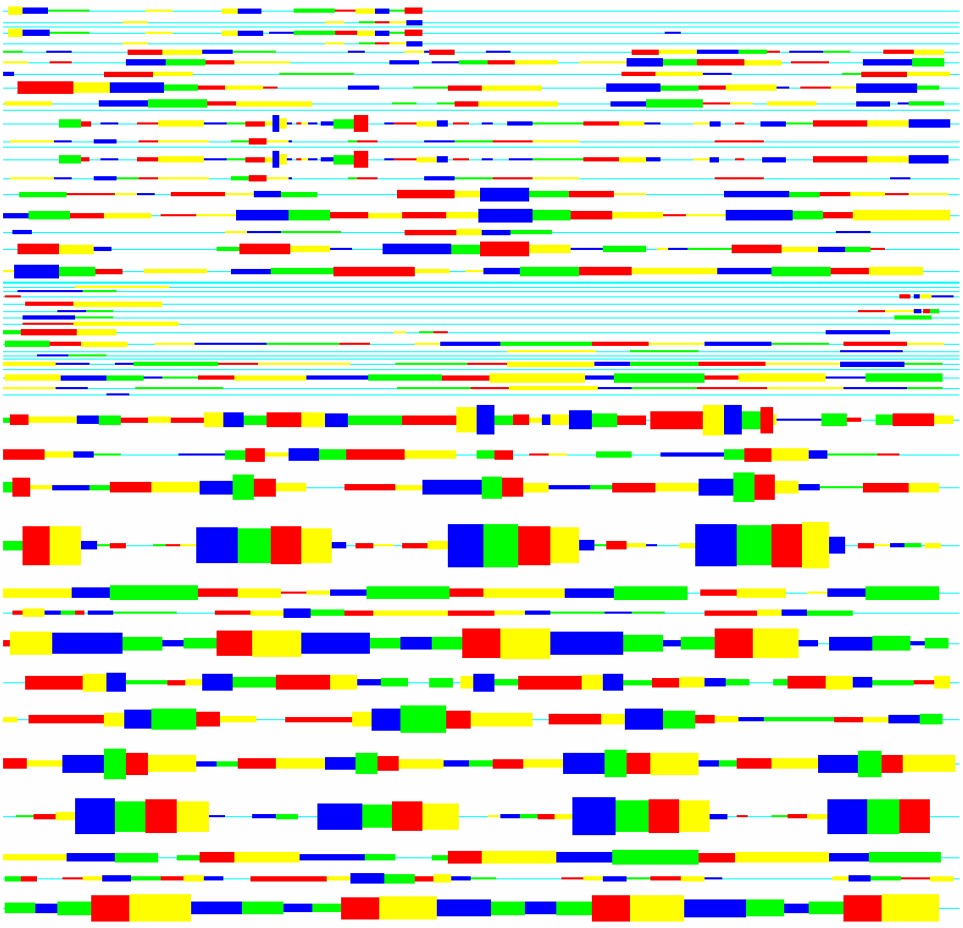
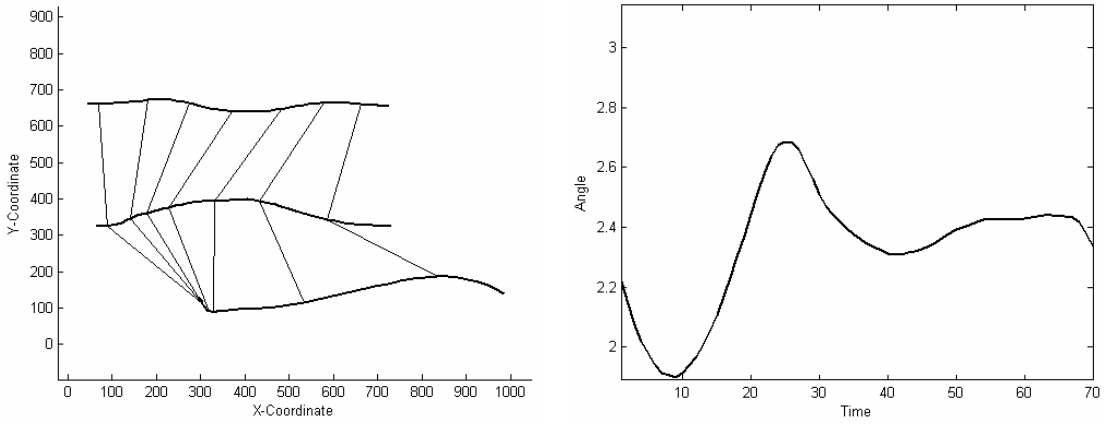


Figure 3.6. Actiongram.

View-invariance

An action representation should be based on primitives robust to variations of the image formation process. View-invariance regards the effect of projecting a 3D representation of human movement into a 2D representation according to a vision system. A view-invariant representation provides the same 2D projected description of an intrinsically 3D action captured from different viewpoints. View-invariance is desired to allow visual perception and motor generation under any geometric configuration in the environment space.



(a) 2D Trajectory. (b) 2D Joint Angle.

Figure 3.7. 2D projected version of the knee joint angle trajectory from a single viewpoint during a walk action.

The view-invariance evaluation requires a 2D-projected version of the initial representative function according to varying viewpoints. For an internal geometric representation, the 3D joint angle is projected according to the two angle sides corresponding to the adjacent body parts, as shown in Figure 3.7. For example, the knee joint 2D angle is formed by the axes of the thigh and shank. These axes are determined by the segments from the hip to the knee joint and from the knee to the

ankle joint. These 3D joints are projected and the 2D joint angle is computed in the projection plane.

To evaluate the view-invariance of our representations, a circular surrounding configuration of viewpoints is used, as shown in Figure 3.8. A viewpoint consists of the camera position (specified by the camera center) and the camera orientation (described by a look-at vector and an upward vector). In our viewpoint configuration, the camera center trajectory corresponds to a circle in 3D space centered at the target point. The look-at vector is oriented from the camera center towards the target point, which is the center of the axis-aligned parallelepiped containing the trajectories of the movement in 3D space. The upward vector has the same orientation as the z-axis vector. The camera center circle is defined as a parametric curve

$$v(\lambda) = [r * \cos(\lambda) + c_x, r * \sin(\lambda) + c_y, c_z],$$

where λ is a parameter representing a direction in degrees from 0° to 360°, r is the radius of the circle, and $[c_x, c_y, c_z]$ is the target point.

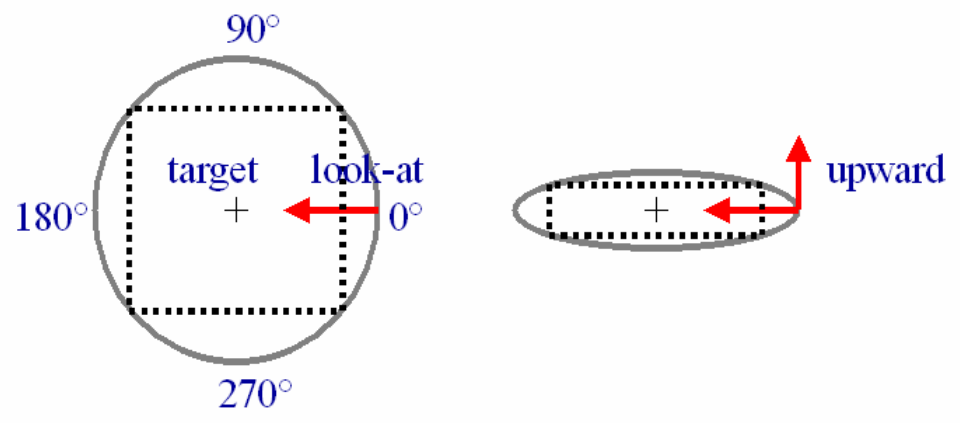


Figure 3.8. A circular configuration of viewpoints.

A view-invariance graph shows for each time instant (horizontal axis) and for each viewpoint in the configuration of viewpoints (vertical axis), the state associated with the movement, as Figure 3.9 shows. A view-invariance measurement concerns the fraction of the most frequent state among all states for all viewpoints at a single instant in time. Let s be a state in our kinetological system, $v_s(t)$ is the fraction of the state s among all viewpoints in our circular configuration at the time instant t. The view-invariance measurement is the maximum value for $v_s(t)$ considering all possible states. A four-state system has a view-invariance measure between 0.25 and 1.0. For each time instant t, the view-invariance measure is computed and plotted on the top of the view-invariance graph. For any joint and any action in our database, the graph demonstrates a high view-invariance measure for our segmentation process, with the only exception at the segment's borders and two degenerated viewpoints.

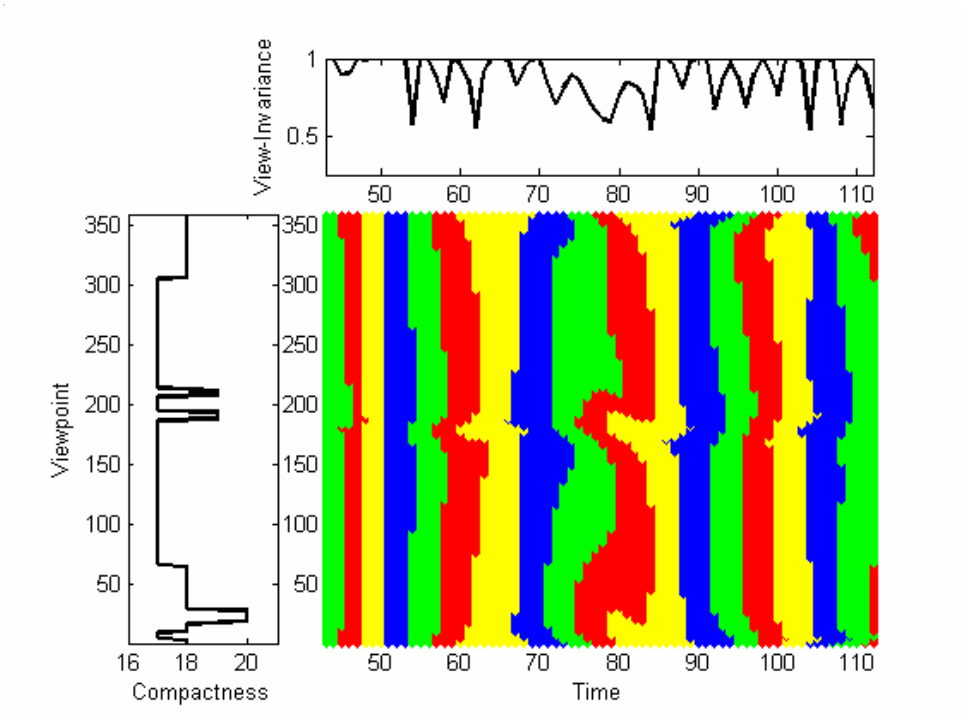


Figure 3.9. View-invariance of the left knee flexion-extension angle during walk.

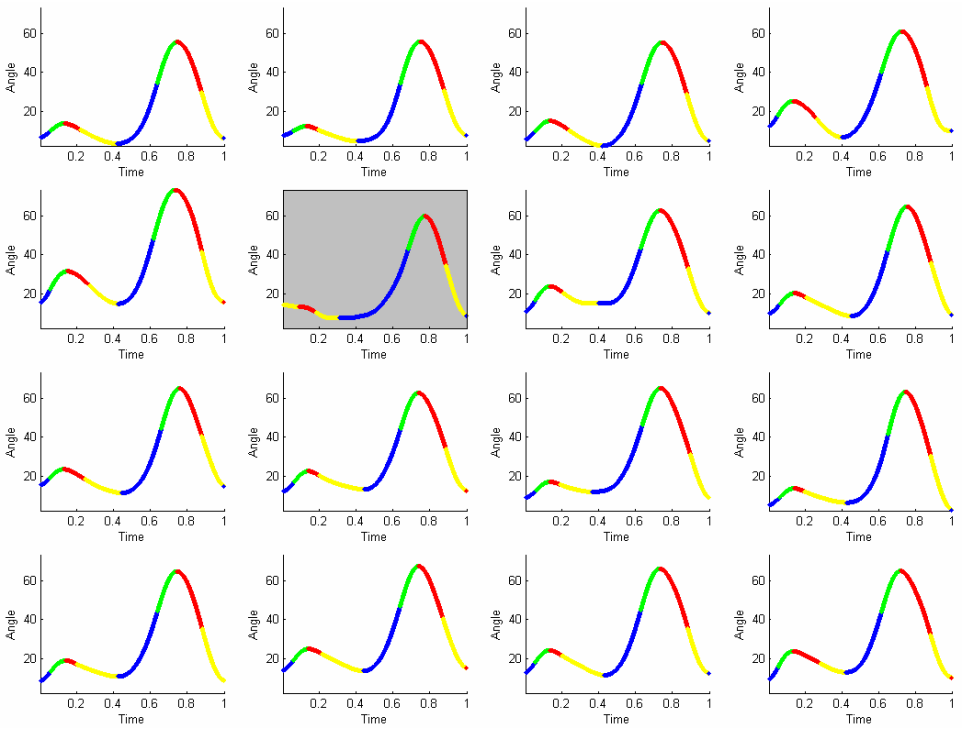
Note that the view-invariance measure has some uncertainty at degenerate viewpoints and at the borders of segments. In these special cases, the movement states are not fully consistent which degrades the view-invariance measure. The degenerate viewpoints are special cases of frontal views where the sides of a joint angle tend to be aligned. In what concerns view-invariance, the border effect shows that movement segments are not completely stable only during the temporal transition between segments. This is analog to coarticulation in speech with similar implications to action recognition tasks.

Reproducibility

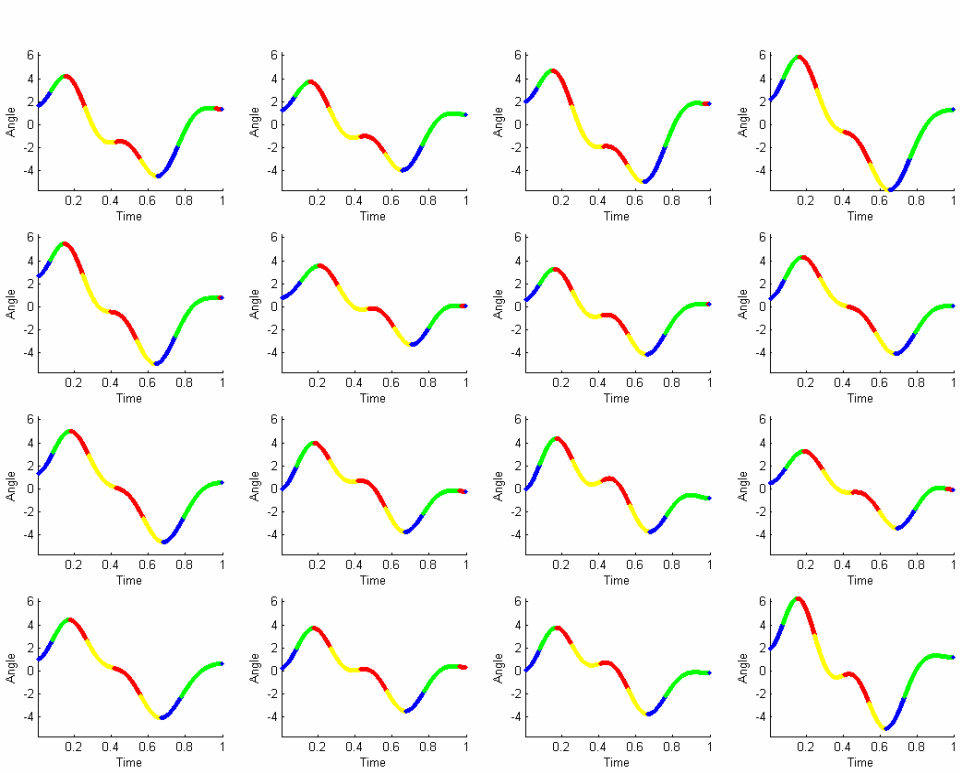
Reproducibility requires an action to have the same description even when a different performance of this action is considered. Intra-personal invariance deals with the same subject performing the same action repeated times. Inter-personal invariance concerns different subjects executing the same action several times. A kinetological system is reproducible when the same symbolic representation is associated with the same action performed at different occasions (intrapersonal) or by different subjects (interpersonal).

To evaluate the reproducibility of our kinetological system, we used human gait data for 16 subjects covering males and females at several ages. For each person, we considered only 12 DOFs associated with the joint angles of the lower limbs: pelvic tilt, pelvic obliquity, pelvic rotation, hip flexion-extension, hip abduction-adduction, hip rotation, knee flexion-extension, knee valgus-varus, knee rotation, ankle dorsi-plantar flexion, foot rotation, and foot progression. A reproducibility measure is computed for each joint angle. The reproducibility measure of a joint angle is the

fraction of the most representative symbolic description among all descriptions for the 16 individuals. A very high reproducibility measure means that symbolic descriptions match among different gait performances and the kinetological system is reproducible. The reproducibility measure is very high for the joint angles which play a primary role in an action, as Figure 3.10 shows for a walking action. The identification of the intrinsic and essential variables of an action is a byproduct of the reproducibility requirement of a kinetological system.



(a) Knee flexion-extension.



(b) Pelvic obliquity.

Figure 3.10. Reproducibility during gait.

Using our kinetological system, six joint angles obtained very high reproducibility: pelvic obliquity, hip flexion-extension, hip abduction-adduction, knee flexion-extension, foot rotation, and foot progression, as shown in Figure 3.11. These variables seem to be the most related to the movement of walking forward. Other joint angles obtained only a high reproducibility measure which is interpreted as a secondary role in the action: pelvic tilt and ankle dorsi-plantar flexion. The remaining joint angles had a poor reproducibility rate and seem not to be correlated to the action but probably to its stability instead: pelvic rotation, hip rotation, knee valgus-varus, and knee rotation. Our kinetological system performance on the reproducibility

measure for all the joint angles shows that the system is reproducible for the DOFs intrinsically related to the action.

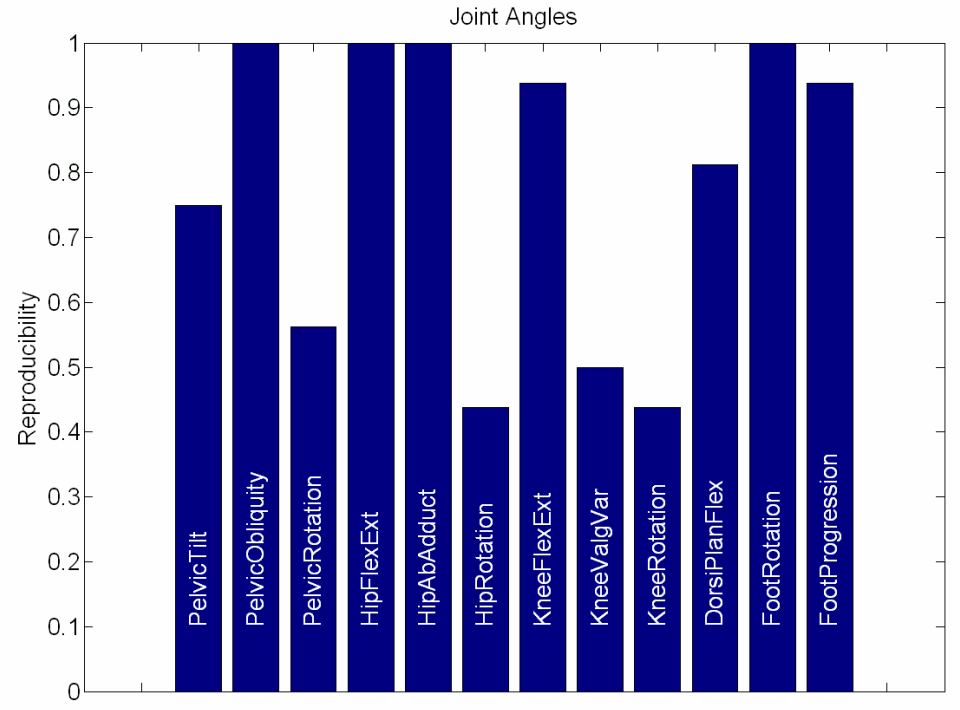
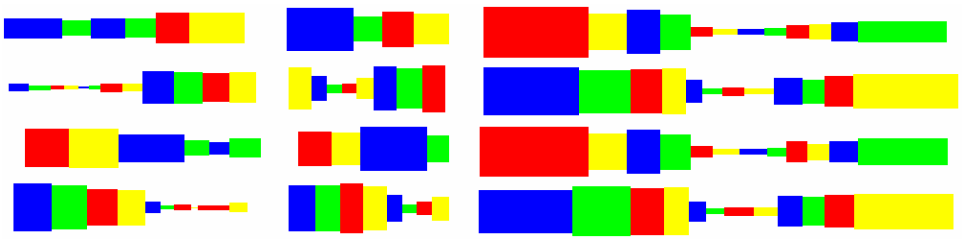


Figure 3.11. Reproducibility measure for 12 DOFs during gait.

Selectivity

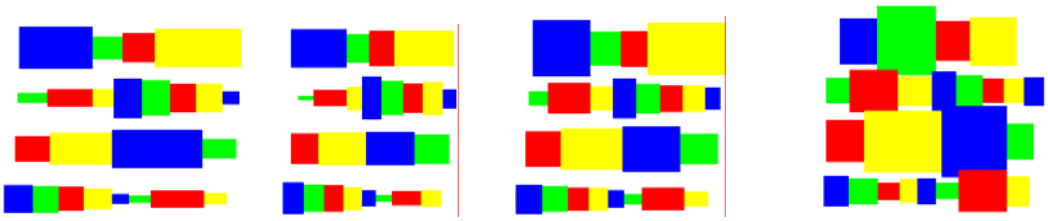
The selectivity principle concerns the ability to discern between distinct actions. In terms of representation, this principle requires a different structure to represent different actions. To evaluate our kinetological system according to the selectivity principle, we compare the compact representation of several different actions and verify whether their structures are dissimilar. The selectivity property is demonstrated using a set of actions performed by the same individual. Four joint angles are considered: left and right hip flexion-extension, left and right knee flexion-extension, as shown in Figure 3.12.



(a) Walk. (b) Run. (c) Jump.

Figure 3.12. Selectivity: Different representations for three distinct actions.

The different actions are clearly represented by different structures. However, manner variations of an action are only different in the quantitative aspect. We investigate the quantitative aspect of four manner variations of the walk action performed by a single subject, as shown in Figure 3.13.



(a) Slow walk. (b) Walk. (c) Walk with stride. (d) Walk with exaggerated stride.

Figure 3.13. Compact representations of four manner variations of the walk action.

Each manner variation has a total of 24 segments for the four joint angles considered. For each pair of manner variations, we compute a dissimilarity vector, where each element corresponds to the difference between the quantitative aspects of the associated segments in the two variations, as shown in Figure 3.14.

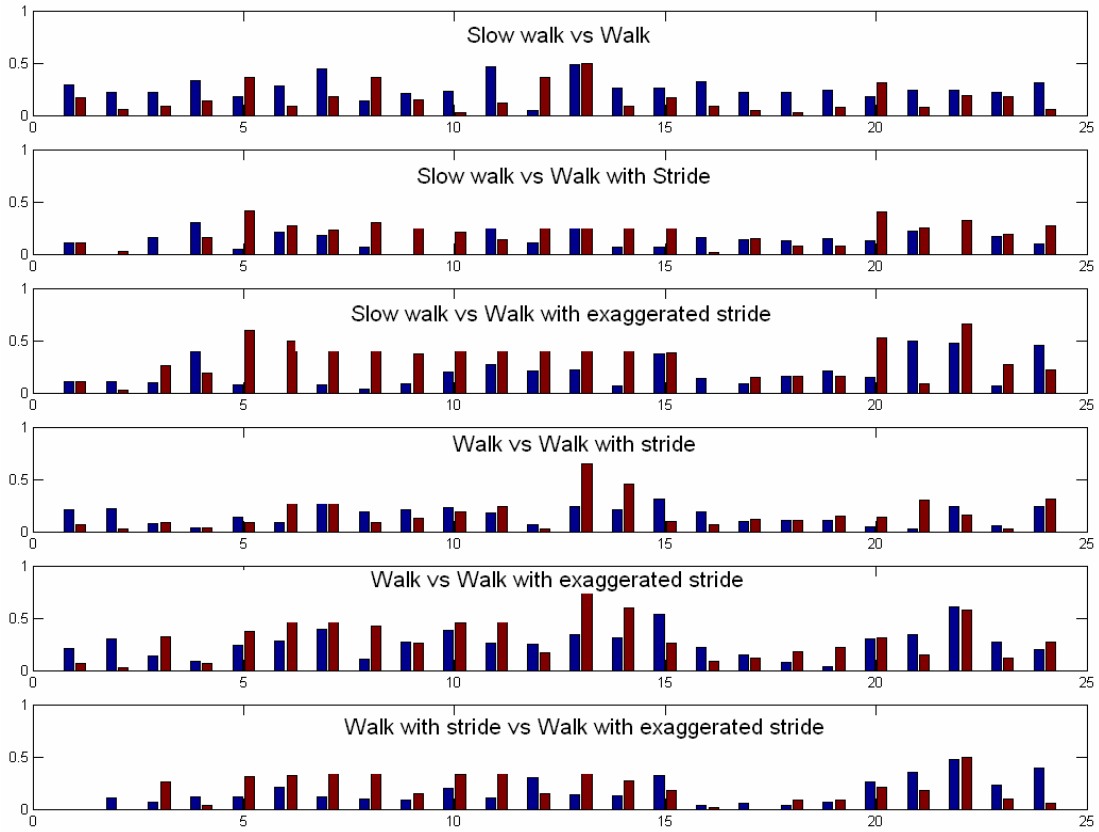


Figure 3.14. Dissimilarity vectors between manner variations of walk: time length (blue) and angular displacement (red).

From these vectors, we can verify the dissimilarity of the manner variations. The closest variations according time length are "Walk with stride" and "Walk with exaggerated stride" (median dissimilarity 12.0%), and according to angular displacement are "Walk" and "Walk with stride" (median dissimilarity 12.2%). This way, even for the same action, the representation has enough dissimilarity to select between different manner variations.

Reconstructivity

Reconstructivity is associated with the ability to reconstruct the original movement signal up to an approximation factor from a compact representation. We propose a

reconstruction method that consists in a novel interpolation algorithm based on the kinetological structure. We consider one segment at a time and concentrate on the state transitions between consecutive segments. Based on a transition, we determine constraints about the derivatives at border points of a segment. Derivatives will have zero value (equation) or a known sign (inequality) at these points.

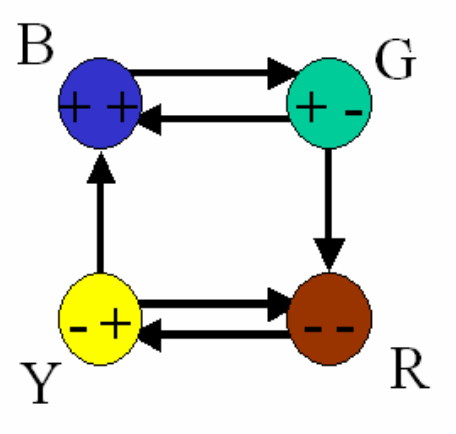


Figure 3.15. Possible state transitions between segments.

For this discussion about reconstructivity, we consider a four-state kinetological system. We investigate the possible state transitions that are feasible in our kinetological system. Each segment can have only two possible states for a next neighbor segment. However, the transition **B** $\rightarrow$ **Y** (**R** $\rightarrow$ **G**) is impossible, since velocity cannot become negative (positive) with positive (negative) acceleration. The kinetological rules of our system are represented by a finite automaton, as shown in Figure 3.15. From these kinetological rules, each of the four segment states has only two possible state configurations for previous and next segments and, consequently, there are eight possible state sequences for three consecutive segments, as shown in Table 3.1. Each possible sequence of three segments corresponds to two equations and two inequality constraints associated with first and second derivatives at border

points t_1 and t_2 of the center segment. Other two inequalities come from the derivatives at interior points ($t_1 < t < t_2$) of the segment.

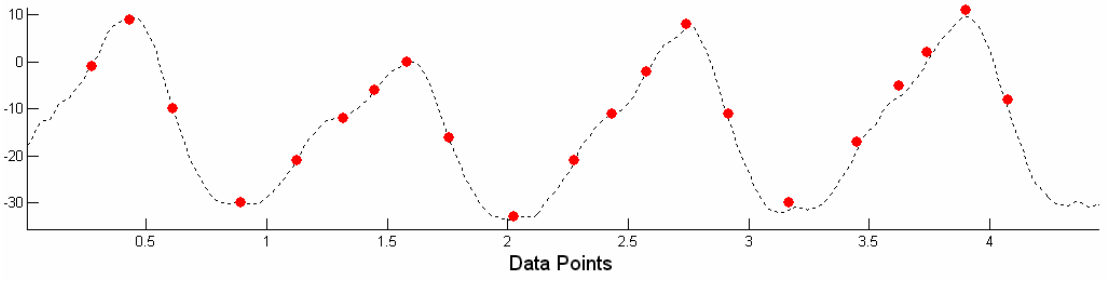
Kinetemes			Border Point t_1		Border Point t_2		Interior Points	
Previous	**Current**	**Next**	$J'(t_1)$	$J''(t_1)$	$J'(t_2)$	$J''(t_2)$	$J'(t)$	$J''(t)$
Y	**B**	**G**	0	+	+	0	+	+
G	**B**	**G**	+	0	+	0	+	+
B	**G**	**R**	+	0	0	-	+	-
B	**G**	**B**	+	0	+	0	+	-
G	**R**	**Y**	0	-	-	0	-	-
Y	**R**	**Y**	-	0	-	0	-	-
R	**Y**	**B**	-	0	0	+	-	+
R	**Y**	**R**	-	0	-	0	-	+

Table 3.1. Possible sequences of neighbor kinetemes and the associated constraints at border points.

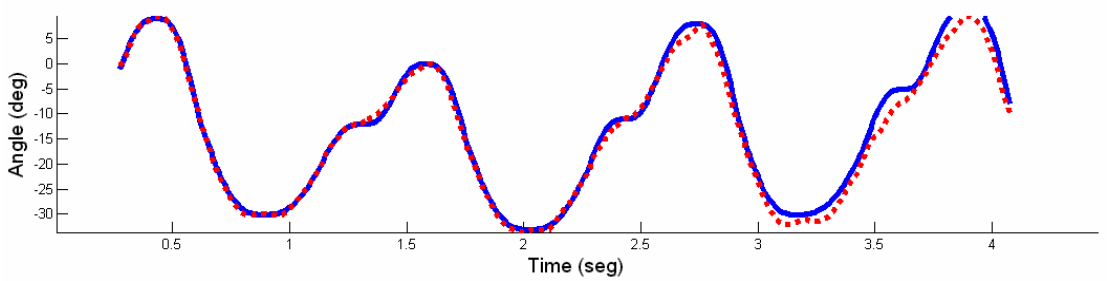
A simple model for the joint angle function during a segment is a polynomial. However, low degree polynomials do not satisfy the constraints originated from the possible sequences of kinetemes. For example, a cubic function has a linear second derivative which is impossible for sequences of segments where the second derivative assumes zero value at the borders and non-zero values at interior points (e.g., a sequence of segments with states **GBG**). The least-degree polynomial satisfying all the constraints imposed by all possible sequences of kinetemes' states is a fourth-degree polynomial.

Using this approach, the reconstruction process needs to find only five parameters defining this polynomial. The polynomial is partially determined with the two associated equations for the particular sequence of kinetemes and two more equations using the joint angle values at the two border points. We obtain these values from the time length and angular displacement of each segment, as shown in Figure 3.16. With four equations, an under-constrained linear system is solved up to one variable. The

last free variable is constrained by four inequalities. This parameter can be determined using some criteria such as jerk (third derivative) minimization.



(a) Sample data points obtained from segments' attributes.



(b) Fourth degree polynomial interpolation considering state transitions.

Figure 3.16. Reconstruction of a joint angle function.

We implemented this reconstruction scheme as a decompression method for motion data, which Figure 3.17 shows. The average error in our motion database was 0.823 degree. Once a reconstruction scheme is provided, the generation of movement from a symbolic representation is feasible. Therefore, the symbolic grammar systems inferred for human actions may be used to effectively generate movement.

Motion Compression and Decompression

An immediate application for a compact representation is compression of motion data. The compression efficiency of our kinetological system was tested in eight different actions extracted from the CMU Motion Capture Library, as shown in Table 3.2. The file size ranges from 128Kb to 1024Kb with increments of 128Kb.

Frame rate is either 120 or 60 frames per second. The median compression rate for all motion files is 3.698%. The best compression is achieved for the actions "answer phone" and "walk" because they consist in smooth movement. Sudden movement, such as the actions "run" and the "Russian dance", obtained the worst performances.

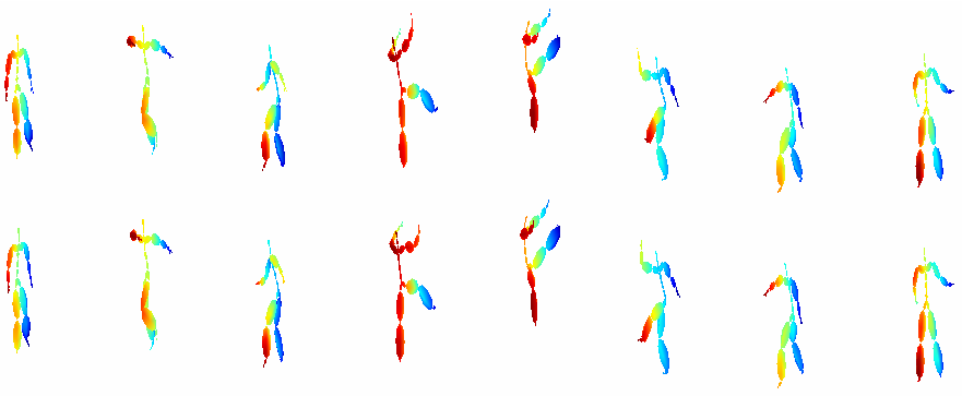


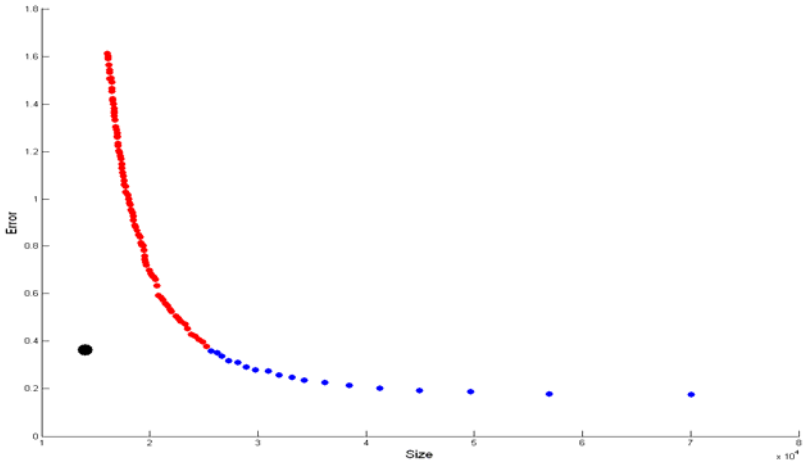
Figure 3.17. Reconstructivity. For the same activity, the top line shows the original motion sequence and the bottom line shows the decompressed one.

Action	Size (Kb)	Frame Rate (frm/sec)	Compression Rate	Average Error (deg)
Run/Jog	128	120	4.075%	0.868
Kick	256	60	3.953%	1.378
Answer Phone	384	60	3.596%	0.363
Walk	512	120	3.625%	0.613
Jump Twist	640	120	3.684%	1.694
Russian Dance	768	120	4.071%	1.804
Weight Lift	896	60	3.657%	0.598
Miscellaneous	1024	120	3.713%	0.778

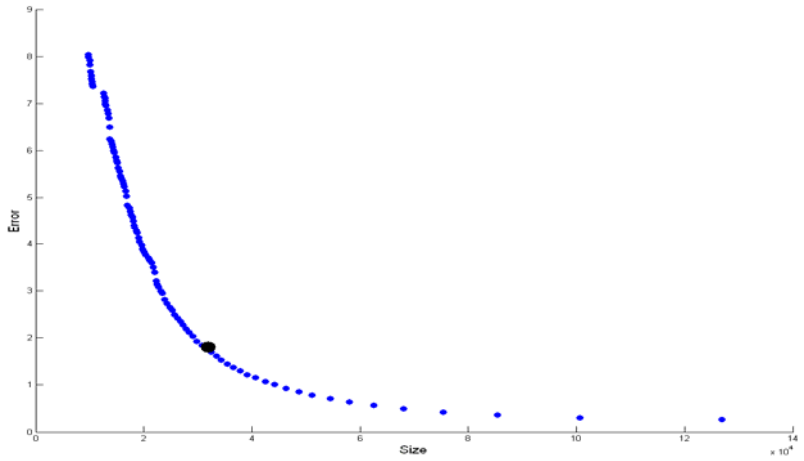
Table 3.2. Experimental motion capture data and results.

The compression encoding is not useful without a decoding process. This way, we implemented the reconstruction proposed and computed the average error of all frames and all joint angles for each action in our test set. The median average error is 0.823 degree. The lowest error is obtained in the "answer phone" action, while the highest error is in the "Russian dance" action. Again, the intuition about smooth and sudden movement takes place to explain the reconstruction errors.

To further evaluate our kinetological system according to compression and decompression, we implemented two other online segmentation methods for comparison purposes. The first method is an online version of the piecewise linear curve simplification [Lim and Thalmann, 2001]. In this algorithm, each segment grows incrementally until the average error is higher than some threshold value. The variation of this threshold parameter leads to a curve of points associated with compression rate and reconstruction error for the algorithm, as shown in Figure 3.18. The single point associated with our method is compared to this curve. In the worst case of the "Russian dance" action, the point stays just above the curve, while in the other actions it is below the curve. This demonstrates that our kinetological system has a better compression rate and reconstruction error performance.



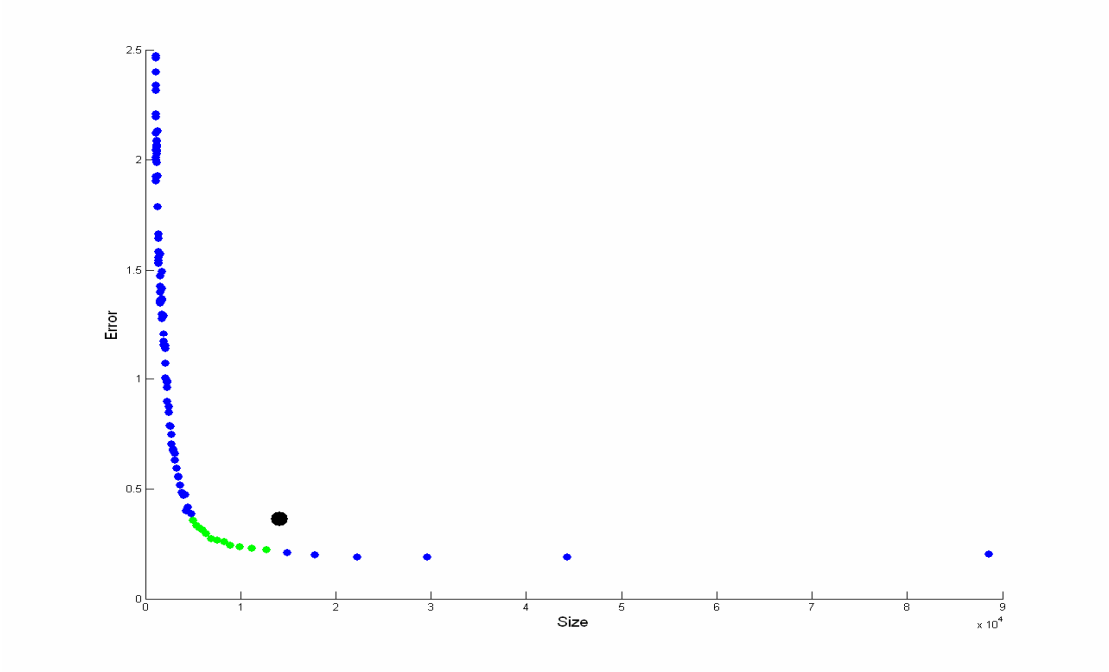
(a) Answer Phone.



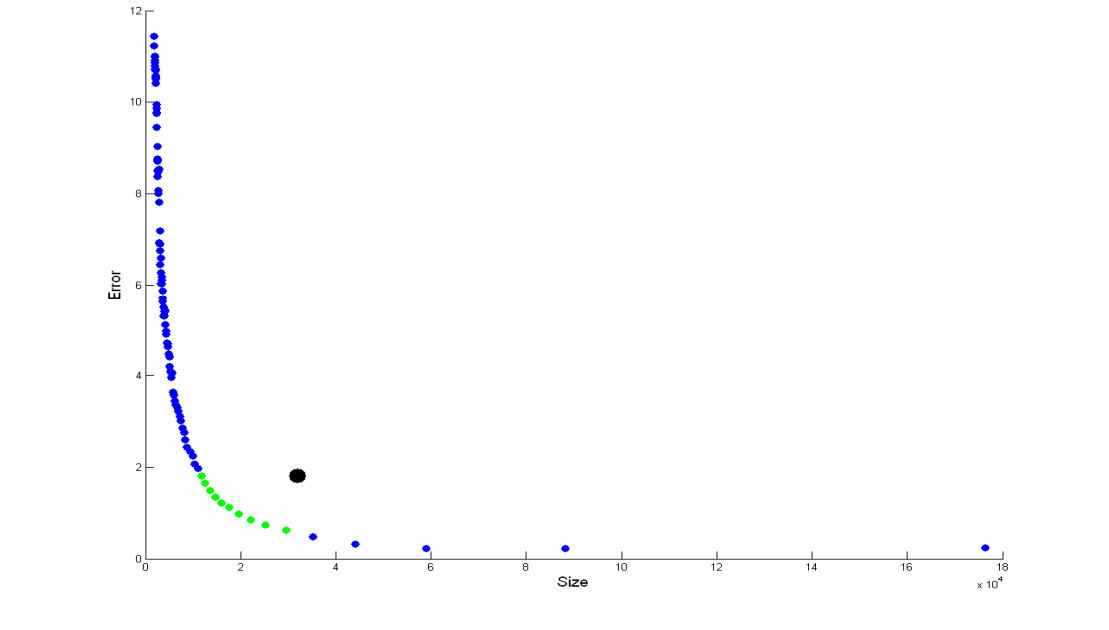
(b) Russian Dance.

Figure 3.18. Compression rate and reconstruction error curve for the piecewise linear method.

The second method implemented for comparison is the uniform sampling [Naka at al., 1999]. In this algorithm, equally spaced frames are selected to represent the motion. The compression rate and reconstruction error curve is computed by varying a parameter that consists in the space between representative frames. Compared to this technique, the single point associated with our method is always above this curve, which shows that our kinetological system has a worse performance, as shown in Figure 3.19. However, the parametric space where the sampling algorithm outperforms our method is limited. Therefore, a search for the best parameter for the sampling algorithm is required to guarantee a better compression than our kinetological system. Furthermore, our method has other applications aimed towards perception and generation of actions.



(a) Answer Phone.



(b) Russian Dance.

Figure 3.19. Compression size and average error curve for the sampling and quantization method.

Conclusion

A kinetological system is the basic structure for an alternative writing/notation system which enables the movement scripting (registration and specification) of human actions. This non-arbitrary symbolic representation provides the means to reason and analyze in terms of movement which enables the understanding of human activities. We believe the importance of a kinetological system to human movement is equivalent to the relevance of a phonological system to spoken language.

Chapter 4: Morphology

Human movement is a natural phenomenon involving a number of independent actuators: articulated body parts or joint angles. The actuators coordinate their actions to achieve some specific common purpose. In human motion modeling literature, the actuators usually consist of a fixed set modeling either total body or a single joint. This assumption neglects the independent behavior of the actuators over different activities. Further, an approach modeling explicitly the variability of the set of actuators is more robust concerning occlusion and field of view limitations in the observation process.

The different strategies of parallel and synchronous interaction among actuators play an important role in human movement. Therefore, a movement representation for a specific human activity should include the set of parallel actuators involved in the activity, the synchronization rules among these actuators, and the motion pattern associated with each participating actuator.

In this chapter, we discuss the morphological part of our linguistic framework where we present the steps required for the construction of a praxicon through the learning of grammar systems for human actions. The discovery of a Human Activity Language involves learning the syntax of human motion which requires the construction of this praxicon. The morphology assumes a non-arbitrary symbolic representation of the human movement. To analyze the morphology of a particular action, we are given a symbolic representation for the motion of each actuator associated with several repeated performances of this action.

This representation originates from kinetology. Movement signals obtained from a motion capture system are divided into consecutive segments according to velocity and acceleration of joint angles. The segments are then transformed into a string of symbols. In fact, symbolization amounts to classifying (clustering) motion segments such that each class contains variations of the same motion.

Given sequences A_i of symbols associated with motor primitives representing the movement for each actuator i when a specific activity is performed repeated times, the problem addressed in this chapter is to identify the set I of essential actuators responsible for the specific goal achieved with this activity, to learn the motion structure for all actuators in I, and the synchronization rules among these actuators. A praxicon is built by solving this problem for all actions in a large lexicon of verbs associated with observable human movement [Guerra Filho and Aloimonos, 2006a]. Although the input concerns a specific action performed several times, we aim to model any general activity, not only restricted to repetitive movement.

We pose this problem as the grammatical inference of a novel grammar system modeling human activity. As a formal model, we propose a Parallel Synchronous Grammar System where each component grammar corresponds to an actuator. We present a novel heuristic parallel learning algorithm to induce this grammar system. Our algorithm does not assume knowledge of either the number of components or the language components of the grammar system being inferred. The input is a single symbolic stream (string) per actuator instead of a sequence of sentences. We evaluated our inference approach with synthetic data and real human motion data. We

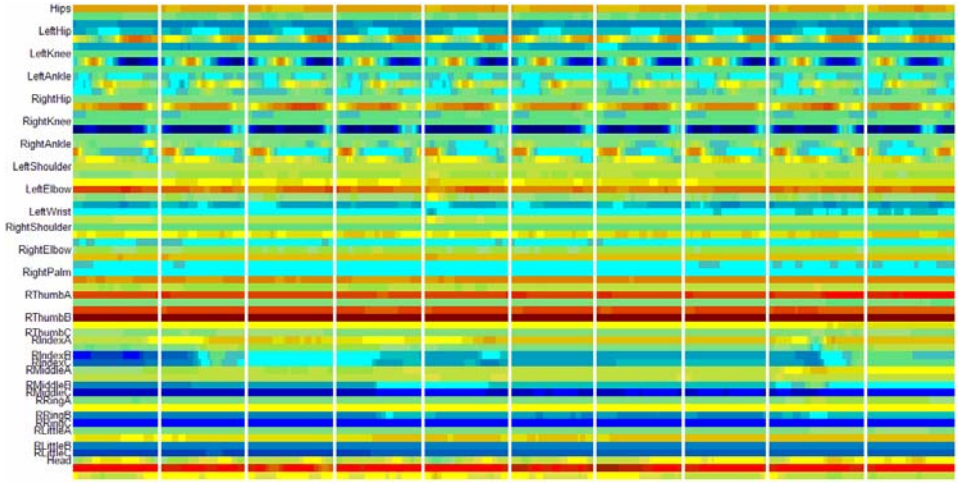
created synthetic actiongrams and tested our method with increasing levels of noise. The algorithm achieved 100% success with a noise level up to 7%.

Morphology is concerned with the structure of words, the constituting parts, and how these parts are aggregated. In the context of a Human Activity Language, morphology involves the structure of each action and the organization of a praxicon in terms of common subparts. Our methodology consists in determining the morphology of each action in a praxicon and then in finding the organization of the praxicon.

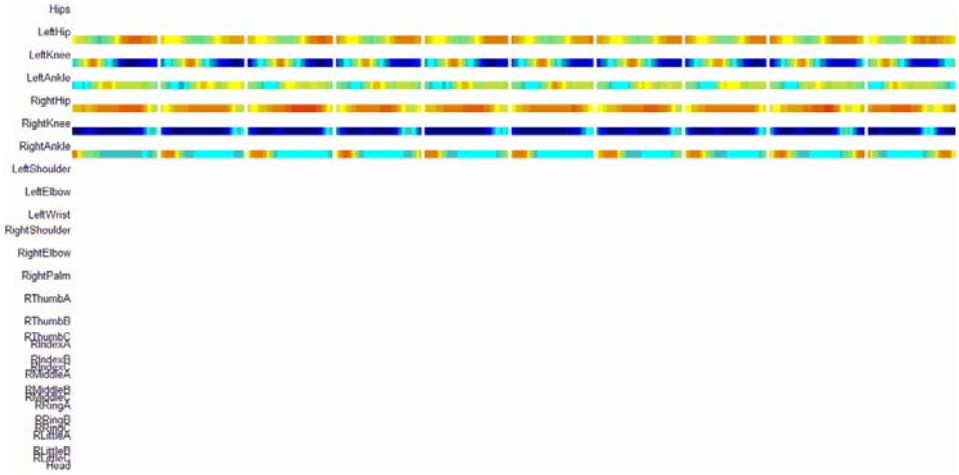
We define a human action morpheme as the set of essential actuators intrinsically involved in the action, the synchronization among these actuators, and the corresponding motion patterns (in terms of kinetemes). The morphemes are the essential parts of human actions. Since the derived motion patterns are sequences of kinetemes, the inference of morphemes is called morpho-kinetology. This part of morphology aims to select a subset of the motion which projects the whole action only into the essential actuators and their motion patterns, as shown in Figure 4.1.

Morpho-Kinetology

The essential actuators are the ones actually responsible for the achievement of the intended result of an action. They are strongly constrained and, consequently, only these "meaningful" actuators will have consistent motion patterns in different performances of the same action. To learn the morphology of a human action, an actiongram associated with several repeated performances of this action is given as input.



(a) The whole motion for all actuators.



(b) The whole motion only for the essential actuators.

(c) Only the motion patterns for the essential actuators.

Figure 4.1. A human action morpheme.

Given such an actiongram A as input, we aim to automatically learn the morpheme of the corresponding action. Formally, the morpheme consists of a set I representing the essential actuators for the action; for each $i \in I$, a substring p_i corresponding to the motion pattern that the actuator i performs during the action; and a set of tuples corresponding to synchronized rules between kinetemes in different strings. Since our input is a set of concurrent strings, we pose this problem as the grammatical inference of a grammar system modeling the human activity such that each component grammar corresponds to an actuator.

Parallel Synchronous Grammar System

In human movement, we are interested only in the simultaneous synchronized work of the components. The communication feature is unnecessary because it is implicit in motion coordination. We propose a novel grammar system, a Parallel Synchronous Grammar System (PSGS), where strings generated by components are not shared through communication steps. The formal model suggested here is based on a PCGS with rule synchronization [Păun, 1993] and no query symbols. The synchronization among rules in different components is modeled as a set of tuples of rules (possibly one rule for each component), where rules in a tuple are derived simultaneously. We specify the definitions related to our adapted PCGS model below. We assume the reader is familiar with the fundamentals of formal language theory. For further information in formal language theory, the reader is directed to [Hopcroft and Ullman, 1979].

A PSGS with $n \geq 1$ components is an $(n+3)$-tuple $\Gamma = (N, T, G_1, G_2, \ldots, G_n, M)$, where N is a set of non-terminals and T is a terminal alphabet (N and T are mutually

disjoint); $G_i = (N, T, P_i, S_i)$, $1 \le i \le n$, are Chomsky grammars with a finite set of production rules P_i over $(N \cup T)$ and a start symbol (axiom) $S_i \in N$; and M is a subset of $(P_1 \cup \{\#\}) \times \ldots \times (P_n \cup \{\#\})$, where $\# \notin (\mathrm{N} \cup \mathrm{T})$ is an additional symbol.

A configuration n-tuple $(x_1, \ldots, x_n)$ of Γ directly derives $(y_1, \ldots, y_n)$, where $x_i, y_i \in (N \cup T)^*$, if we have a direct derivation $x_i \Rightarrow y_i$ in each grammar G_i with x_i not terminal or $x_i = y_i$ when $x_i \in T^*$. Each component uses one of its rewriting rules except those grammars which have already produced a terminal string. At a derivation step, a transition n-tuple $(p_1, \ldots, p_n)$ of M is applied, that is $x_i \Rightarrow y_i$ by the rule p_i, if $p_i \in P_i$, and $x_i = y_i$, if $p_i = \#$. A derivation starts from the initial configuration consisting of the axioms $(S_1, \ldots, S_n)$. The language generated by Γ is

$$L(\Gamma) = \{(\alpha_1, \ldots, \alpha_n),\ \alpha_i \in T^* \mid (S_1, \ldots, S_n) \Rightarrow^* (\alpha_1, \ldots, \alpha_n)\}.$$

A simple example of a PSGS with four components is

$\Gamma = (\{S_1, S_2, S_3, S_4, N_1, \ldots, N_{23}\}, \{a, b, c, d\}, G_1, G_2, G_3, G_4, M)$, where

$P_1 = \{S_1 \to N_{13}S_1,\ S_1 \to N_{13},\ N_5 \to \mathrm{bc},\ N_9 \to \mathrm{a}N_5,\ N_{10} \to N_9\mathrm{d},\ N_{11} \to N_{10}N_5,$
$N_{12} \to N_{11}\mathrm{a},\ N_{13} \to N_{12}\mathrm{d}\}$,

$P_2 = \{S_2 \to N_{18}S_2,\ S_2 \to N_{18},\ N_1 \to \mathrm{bc},\ N_{14} \to N_1\mathrm{a},\ N_{15} \to N_{14}\mathrm{d},\ N_{16} \to N_{15}\mathrm{a},$
$N_{17} \to N_{16}N_1,\ N_{18} \to N_{17}\mathrm{d}\}$,

$P_3 = \{S_3 \to N_7S_3,\ S_3 \to N_7,\ N_2 \to \mathrm{cd},\ N_3 \to N_2\mathrm{a},\ N_4 \to N_3\mathrm{b},\ N_7 \to N_4N_4\}$,

$P_4 = \{S_4 \to N_{23}S_4,\ S_4 \to N_{23},\ N_6 \to \mathrm{bc},\ N_{17} \to \mathrm{a}N_6,\ N_{20} \to N_{19}\mathrm{d},\ N_{21} \to N_{20}N_6,$
$N_{22} \to N_{21}\mathrm{a},\ N_{23} \to N_{22}\mathrm{d}\}$, and

$\mathrm{M} = \{(S_1 \to N_{13}S_1,\ S_2 \to N_{18}S_2,\ S_3 \to N_7S_3,\ S_4 \to N_{23}S_4),\ (S_1 \to N_{13},\ S_2 \to N_{18},$
$S_3 \to N_7,\ S_4 \to N_{23}),\ (N_5 \to \mathrm{bc},\ N_1 \to \mathrm{bc},\ N_4 \to N_3\mathrm{b},\ N_6 \to \mathrm{bc}),\ (N_9 \to \mathrm{a}N_5,$

$N_{14} \rightarrow N_1$a, #, $N_{19} \rightarrow$ aN_6), ($N_{10} \rightarrow N_9$d, $N_{15} \rightarrow N_{14}$d, #, $N_{20} \rightarrow N_{19}$d), ($N_{11} \rightarrow N_{10}N_5$, $N_{16} \rightarrow N_{15}$a, #, $N_{21} \rightarrow N_{20}N_6$), ($N_{12} \rightarrow N_{11}$a, $N_{17} \rightarrow N_{16}N_1$, #, $N_{22} \rightarrow N_{21}$a), ($N_{13} \rightarrow N_{12}$d, $N_{18} \rightarrow N_{17}$d, $N_7 \rightarrow N_4N_4$, $N_{23} \rightarrow N_{22}$d)}.

An example derivation in Γ is (S_1, S_2, S_3, S_4) $\Rightarrow$ (N_{13}, N_{18}, N_7, N_{23}) $\Rightarrow$ (N_{12}d, N_{17}d, N_4N_4, N_{22}d) $\Rightarrow$ (N_{11}ad, $N_{16}N_1$d, N_4N_4, N_{21}ad) $\Rightarrow$ ($N_{10}N_5$ad, N_{15}aN_1d, N_4N_4, $N_{20}N_6$ad) $\Rightarrow$ (N_9dN_5ad, N_{14}daN_1d, N_4N_4, N_{19}dN_6ad) $\Rightarrow$ (aN_5dN_5ad, N_1adaN_1d, N_4N_4, aN_6dN_6ad) $\Rightarrow$ (abcdbcad, bcadabcd, N_3bN_3b, abcdbcad) $\Rightarrow$ (abcdbcad, bcadabcd, N_2abN_2ab, abcdbcad) $\Rightarrow$ (abcdbcad, bcadabcd, cdabcdab, abcdbcad). The corresponding parse trees displaying the structure of this set of strings are shown in Figure 4.2.

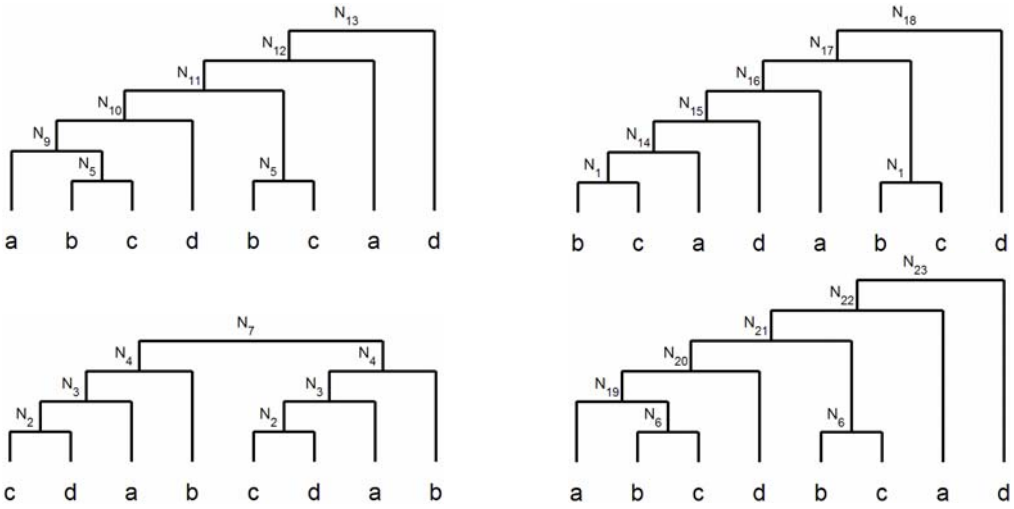


Figure 4.2. Parse trees for a Parallel Synchronous Grammar System.

A PSGS consists in a set of CFGs related by synchronized rules. This grammar models a system with a set A of different concurrent strings A_i: an actiongram. Each string A_i in an actiongram corresponds to the language which will be inferred for a component grammar G_i modeling an actuator. Each symbol $A_i(j)$ in a string corresponds to a pair ($T_i(j)$, $D_i(j)$) for $i = 1, \ldots, m_i$. $T_i(j)$ is the start time and $D_i(j)$ is the

time length of the segment corresponding to $A_i(j)$. Note that $A_i(j) \neq A_i(j+1)$ and $T_i(j) + D_i(j) = T_i(j+1)$.

Parallel Learning

The execution of a human action involves the achievement of some goal and, therefore, requires consistency in a single string (sequential grammar) and coordination among different strings (parallel grammar). This way, sequential grammar learning and parallel grammar learning are combined to infer the morphology of a human action.

We propose parallel learning to concurrently infer a grammar system as the structure of all strings $A_1, \ldots, A_n$ in the actiongram A. Our Parallel Learning (PAL) algorithm executes the sequential learning within each string A_i independently, as shown in Figure 4.3. The digram frequency is still computed within the string corresponding to each joint angle independently. The function *DigramFrequency* finds a matrix *df*, where each element $df(i, j)$ is the number of occurrences of digram $A_i(j)A_i(j+1)$ in string A_i. A new rule is created for the digram $A_i(j)A_i(j+1)$ corresponding to element (i, j) with the current maximum frequency in matrix *df*. A non-terminal N_c corresponding to a rule $[N_c \rightarrow A_i(j)\ A_i(j+1)]$ is inserted in the set of rules P_i. The procedure *ReverseRewrite* replaces each occurrence of the digram $A_i(j)A_i(j+1)$ in string A_i with the non-terminal N_c. A new non-terminal is associated with the interval corresponding to the union of time intervals of both symbols $A_i(j)$ and $A_i(j+1)$ in the digram. In parallel learning, nodes are merged only if the new rule is synchronized with other rules in different CFG components of a grammar system. This way,

overgeneralization is avoided since synchronization guarantees a relation between the merged rules.

```
Algorithm PAL(A, T, D)
  df ← DigramFrequency(A);
  while (∃i | m_i > 1 and max(df) > 1)
    (i, j) ← argmax(df);
    P_i ← P_i ∪ [N_c → A_i(j) A_i(j+1)];
    ReverseRewrite(A, c, i, j);
    R ← SynchronizedRules(A,T,D,R,c,i);
    df ← DigramFrequency(A);
  end
```

```
Function SynchronizedRules(A,T,D,R,c,i)
  E_c ← FindOccurrences(A_i, N_c);
  for k = 1, …, c-1
    if (i ≠ q, where N_c ∈ A_i and N_k ∈ A_q)
      E_k ← FindOccurrences(A_q, N_k);
      for u = 1, …, |E_c|; v = 1, …, |E_k|
        if (E_c(u) ∩ E_k(v))
          I(u, v) ← 1;
        end
      end
      if (one-to-one(I))
        R ← R ∪ (N_c, N_k);
      end
    end
  end
```

Figure 4.3. Parallel Learning algorithm.

Each new non-terminal N_c is checked for possible synchronized rules with existing non-terminals in the CFGs of other strings ($i \neq q$), as shown in Figure 4.4. Synchronization between two non-terminals (N_c and N_k) in different CFGs requires each occurrence of these non-terminals (obtained with procedure *FindOccurrences*) to have intersecting time intervals ($E_c(u) \cap E_k(v)$) in the different strings generated by their respective CFGs. Synchronization relating two non-terminals in different CFGs is issued if there is a one-to-one mapping (*one-to-one*(I)) of their occurrences in the associated strings. Further, any two mapped occurrences must correspond to intersecting time periods. The function *SynchronizedRules* performs this search for synchronization and incrementally creates a relation R, where each pair in this relation represents two synchronized rules in different component grammars. The synchronous tuples in M are recovered from R.

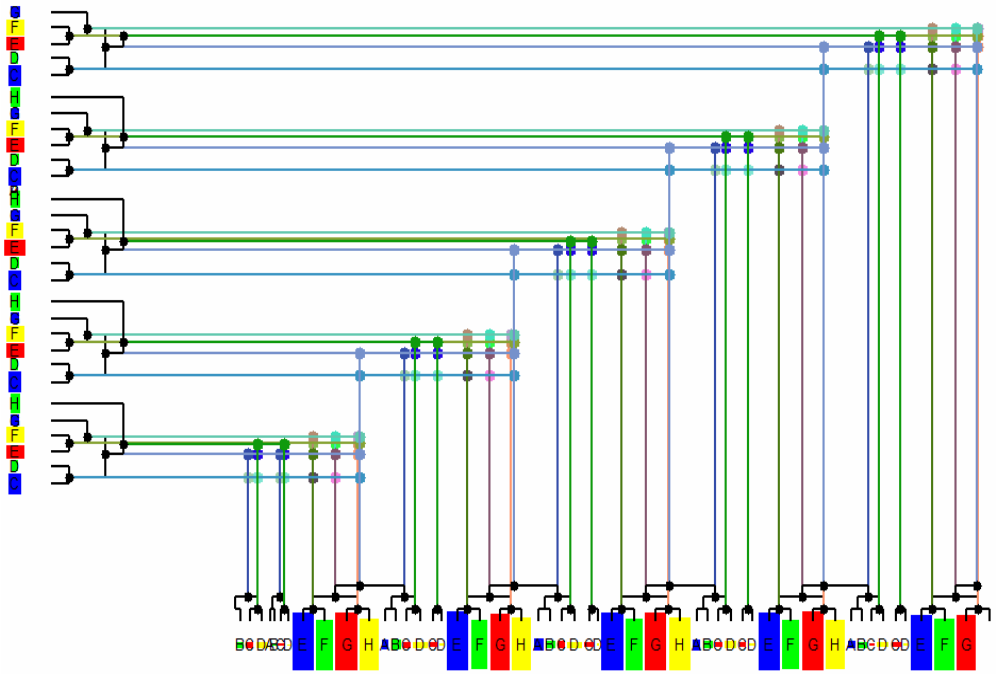
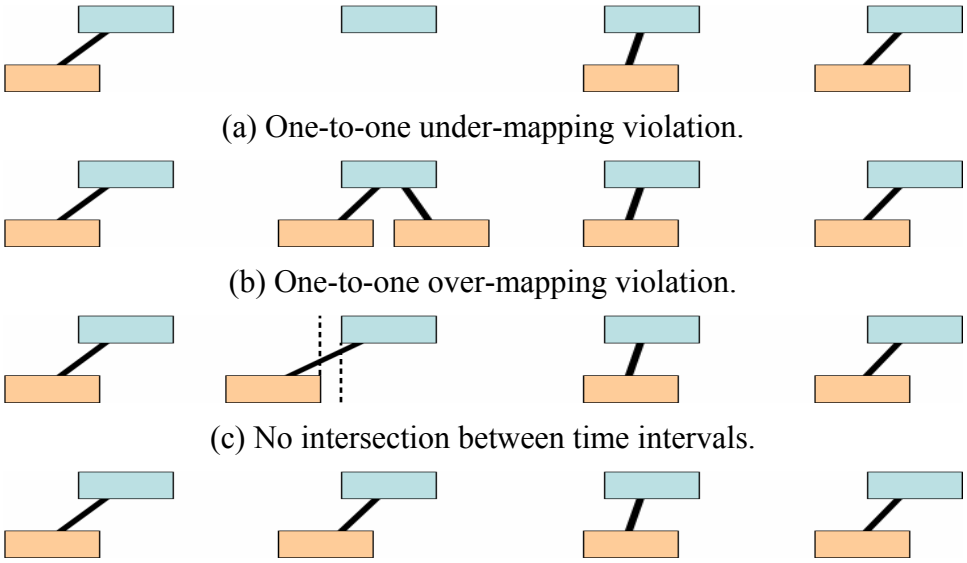


Figure 4.4. Two CFGs (corresponding to hip and knee flexion-extension) related by synchronized rules of a PSGS.

Figure 4.5 illustrates the constraints for synchronized rules. We show two non-terminals in different CFGs represented as rectangles with two different colors. These non-terminals are displayed in different rows such that each rectangle corresponds to one occurrence of the non-terminal. The horizontal position and length of each occurrence illustrates the respective time interval.

We show an execution of our parallel algorithm below. For two iterations, we show the set of strings A, the sets of production rules P_i, and the relation R with the synchronized rules. The input set of strings is derived from the previous example of PSGS with an additional spurious string: A_4. Dashes are used just for visual presentation of the time period associated with each symbol in A. Non-terminals are displayed only with their index numbers.



(a) One-to-one under-mapping violation.

(b) One-to-one over-mapping violation.

(c) No intersection between time intervals.

(d) All constraints satisfied.

Figure 4.5. Constraints for synchronized rules.

```
A = {(a-5d-5ada-5d-5ada-5d-5ad),
     (-1ada-1d-1ada-1d-1ada-1d),
     (--4---4---4---4---4---4-),
     (adadcabcadbbdbcacdcbbaad),
     (a-6d-6ada-6d-6ada-6d-6ad)},
P1 = {5 -> bc},
P2 = {1 -> bc},
P3 = {2 -> cd, 3 -> 2a, 4 -> 3b},
P4 = {},
P5 = {6 -> bc},
R  = {(2, 1), (3, 1), (4, 1), (5, 1), (5, 2), (5, 3),
     (5, 4), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)}.

A = {(a-5d-5ada-5d-5ada-5d-5ad),
     (-1ada-1d-1ada-1d-1ada-1d),
     (----7-------7-------7---),
     (adadcabcadbbdbcacdcbbaad),
     (a-6d-6ada-6d-6ada-6d-6ad)},
P1 = {5 -> bc},
P2 = {1 -> bc},
P3 = {2 -> cd, 3 -> 2a, 4 -> 3b, 7 -> 44},
P4 = {},
P5 = {6 -> bc},
R  = {(2, 1), (3, 1), (4, 1), (5, 1), (5, 2), (5, 3),
     (5, 4), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5)}.
```

In practice, synchronization is difficult to be detected for low-level non-terminals (closer to the leaves of the grammar tree forest). These non-terminals have a high frequency and some atom occurrences are spurious. However, high-level non-terminals are more robust and synchronization is reliably detected for them. To overcome this problem, the algorithm could be adapted with a re-check for synchronization. When synchronization is issued for a pair of non-terminals *A* and *B*, their descendents in the respective grammar trees are re-checked for synchronized rules. This time, we consider only instances of their descendent non-terminals which are concurrent with *A* and *B*, respectively.

Besides formally specifying the relations between CFGs, the synchronized rules are effective in identifying the maximum level of generalization for an action as demonstrated with the non-terminal `7` above. Further, the set of strings related by synchronized rules corresponds to the actual grammar components. The basic idea is to eliminate non-terminals with no associated synchronization and the resulting grammars are the true components of the learned PSGS. Note that the grammar associated with string A_4 above will end up with three non-synchronized rules (`P4 = {8 -> ad, 28 -> ca, 29 -> bb}`), which correctly identifies it as the spurious string not belonging to the grammar system inferred.

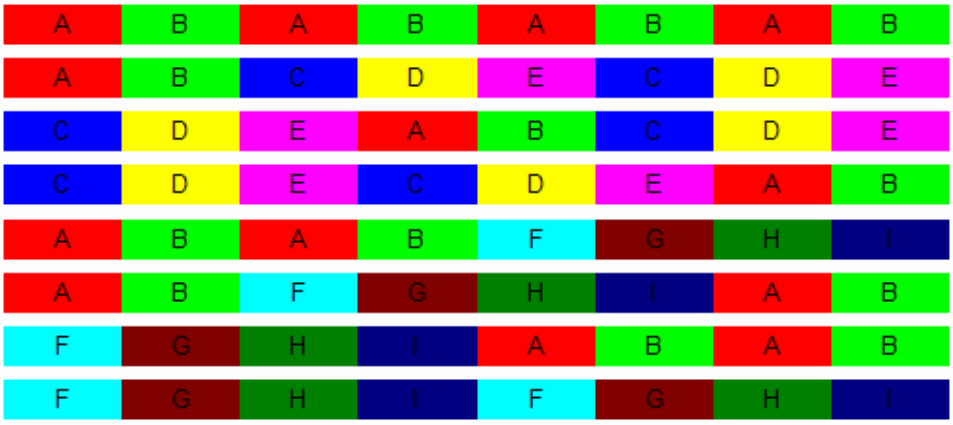
To identify the essential actuators and the corresponding motion patterns, the non-terminals associated with no synchronization rules are discarded from the component CFGs. The set *I* of essential actuators is identified according to the set of CFGs with a considerable amount of synchronized rules. For each actuator $i \in I$, the associated

motion pattern p_i is generated by the non-terminal in G_i whose occurrences cover the most time of the duration of the motion.

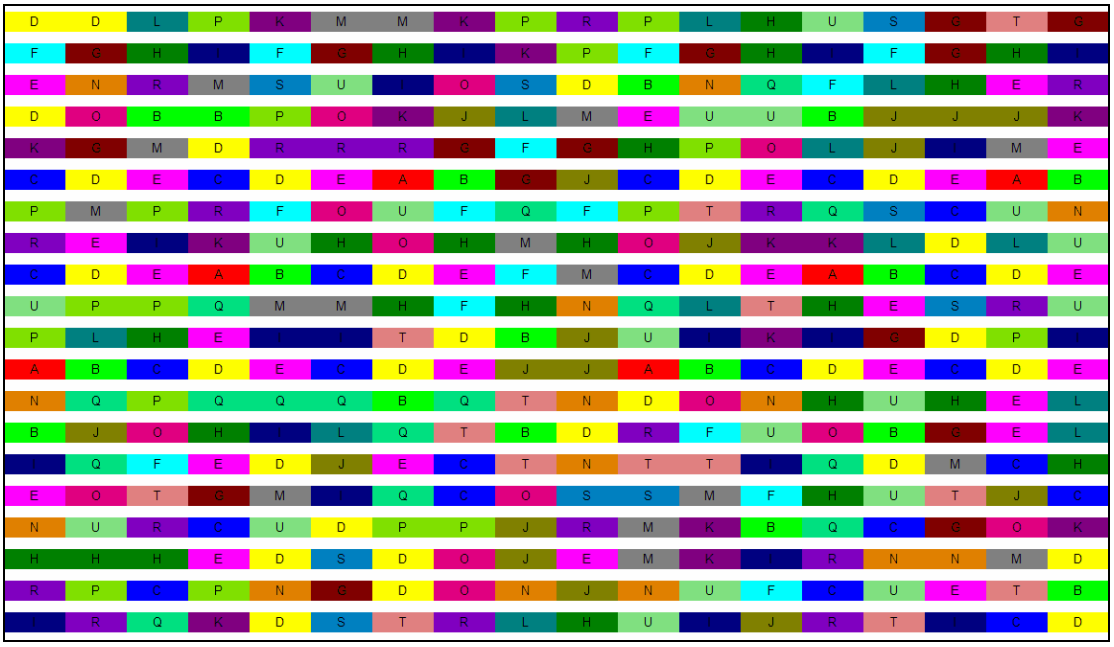
Using the synchronized rules, we prune spurious production rules in the component grammars. Consequently, the remaining rules serve to identify the subset of true components related to the action. The resulting component grammars correspond to the actuators coordinated for the achievement of a common purpose embedded in the action. Overgeneralized rules are also discarded due to the lack of synchronization. Therefore, the remaining highest-level in each grammar component delimits the motion pattern associated with the action.

Evaluation

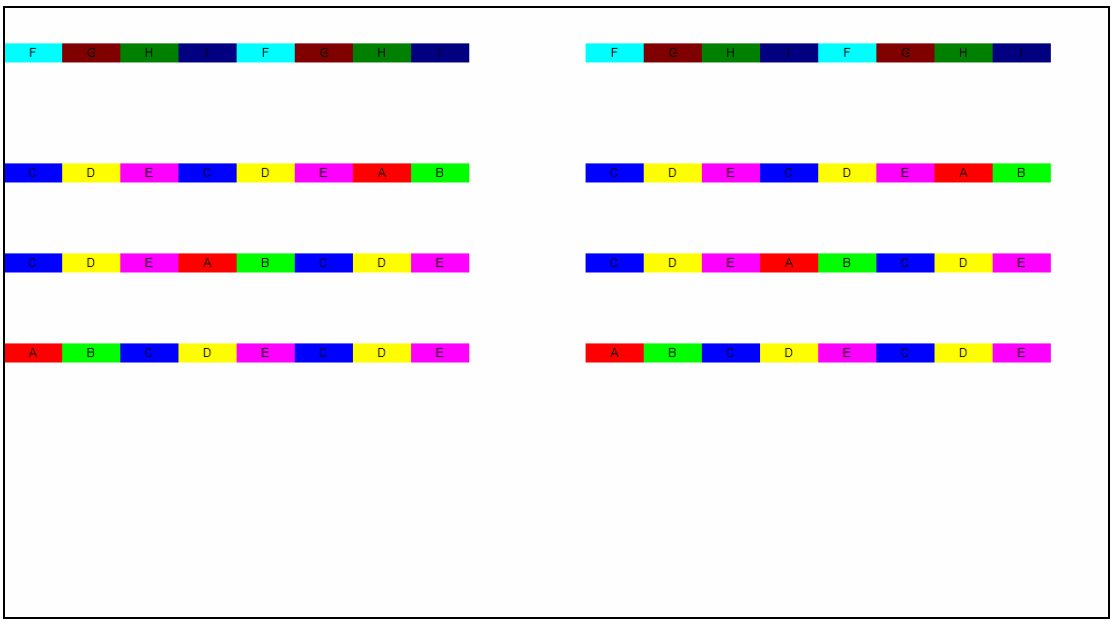
We evaluated our parallel algorithm with synthetic data and real human motion data. A synthetic actiongram was created with 20 synchronous strings, each one containing 100 segments with a uniform time length. Each segment is associated with a symbol extracted from an alphabet of 20 characters. Four synchronous strings in the actiongram are created according to a pattern chosen among one of eight different templates, as shown in Figure 4.6a. These templates are repeated 10 times along the patterned string (separated by two random characters) to represent a consistent movement performed several times. Different templates are applied to the four patterned strings synchronously. The remaining strings are generated with random symbols from the alphabet to simulate spurious movement, shown in Figure 4.6b.



(a) Pattern templates.



(b) Synthetic actiongram.



(c) Ground truth.

Figure 4.6. Evaluation with synthetic data.

The ground truth for our problem is available in a synthetic actiongram, as shown in Figure 4.6c. We compare the output of our algorithm with this ground truth to define an evaluation criterion. If the output matches the ground truth, i.e., all four pattern strings are identified and the corresponding templates are extracted, we claim that the algorithm was successful.

For a more realistic evaluation, we inserted noise in the synthetic data. The four patterned strings have a number of symbols replaced by noisy random characters in the alphabet. We tested our algorithm 100 times for an increasing level of noise and computed the overall success rate for each noise level, as shown in Figure 4.7. The algorithm achieves 100% success rate up to 7% of noise inserted in the patterned strings. The algorithm is robust even at 10% of noise level when the success rate was 96%.

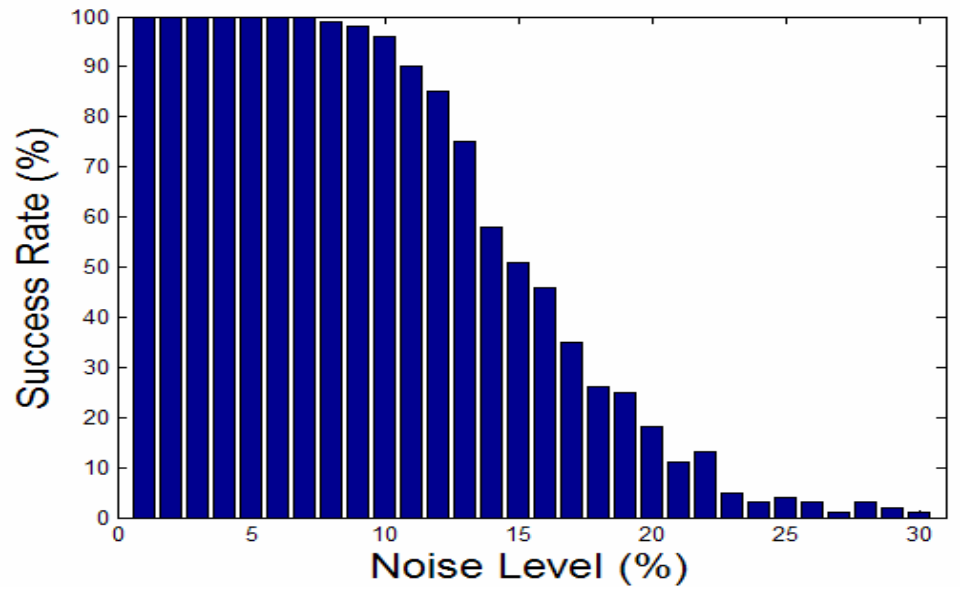


Figure 4.7. Evaluation with increasing noise levels.

Action Morphology Inference

Given an actiongram of a real human activity, parallel learning selects a subset of the actiongram which projects the whole action only into the intrinsic joint angles and motion patterns of the action. The whole grammatical inference process is data driven. We validated our approach with a large scale motion capture database. This process of morpheme learning was performed in each action of our motion database (see Appendix B). Our database consists of about 200 actions associated with English verbs related to observable voluntary meaningful movement. Our database does not consist of actions in any specific domain; instead it contains general activities covering locomotion, non-locomotion, manipulative, and interactive actions. The subset of induced grammar components is associated with joint angles concerned intrinsically with the action. The resulting grammars represent the morphological structure of the action being induced. We automatically identified the morphemes in our database, i.e., the essential actuators participating in each action, the associated

motion patterns (described as sequences of kinetemes), and their synchronization with movement in other joints. In Figure 4.8, we display the motion patterns for the "right hip flexion-extension" actuator in the actions of our database which this actuator is an essential actuator.

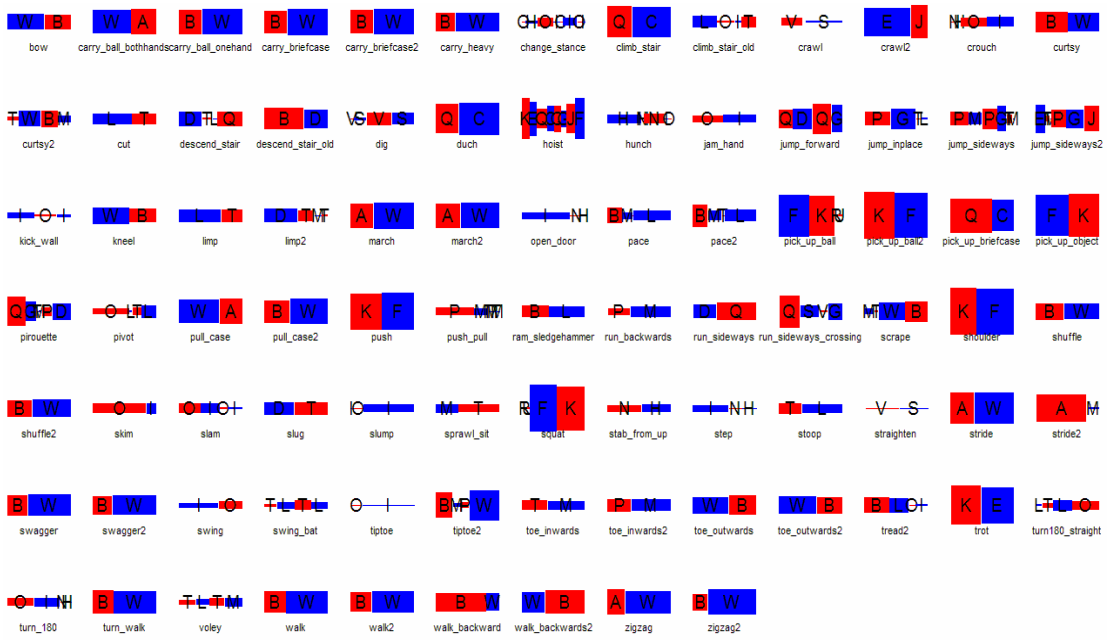


Figure 4.8. The "right hip flexion-extension" motion patterns.

Motion patterns of different actions for a particular actuator may have a common structure. Some motion patterns share the same kineteme depicted as segments with the same symbol in Figure 4.8. This way, the morphological grammars become even more compact with just a few kinetemes required to represent all motions.

Morpho-Syntax

Once morphemes are inferred for each action in a praxicon, we may learn further structure for these morphemes. This structure arises from the ordering, intersection,

and repeated occurrences of kinetemes in motion patterns for the same actuator but in different actions. We refer to this additional structure as morpho-syntax.

Our method to infer morpho-syntax considers a single actuator i at a time. We denote p_i^a as the motion pattern for actuator i and action a, such that $i \in I^a$, where I^a is the set of essential actuators for action a. Basically, all motion patterns p_i^a for actuator i in different actions are described as sequences of kinetemes. These sequences altogether can be generated by a single context-free grammar that represents a more compact and efficient structure: a morphological grammar.

Initially, the symbolization process is performed considering the segments associated with kinetemes in motion patterns p_i^a for all actions. This way, segments of different actions may become associated with the same symbol. In other words, the same kineteme or motor primitive may be found in different actions. With regards to actuator i, this symbolization results in a set of symbols that represents a unified alphabet of kinetemes for all actions in the praxicon. The motion patterns for actuator i in all actions are rewritten according to this unified alphabet. In our experiments, for a total of 30 actuators evaluated, the maximum size of such an alphabet was 31 kinetemes and the median size was 17 kinetemes.

Overlapping kinetemes in joint angle space are considered different units without taking their angular intersection in consideration. To overcome this lack of structure, we subdivide the original kinetemes according to their intersections with other kinetemes in joint angle space. In this space, a kineteme ranges from an initial angle to a final angle. In Figure 4.9, each kineteme is represented by a rectangle where the left side is at the initial angle and the right side is at the final angle. These angles are

displayed as vertical lines. The border angles correspond to points where the kinetemes are subdivided. Therefore, the intervals delimited by these angles correspond to new kinetemes (shown as red symbols in Figure 4.9).

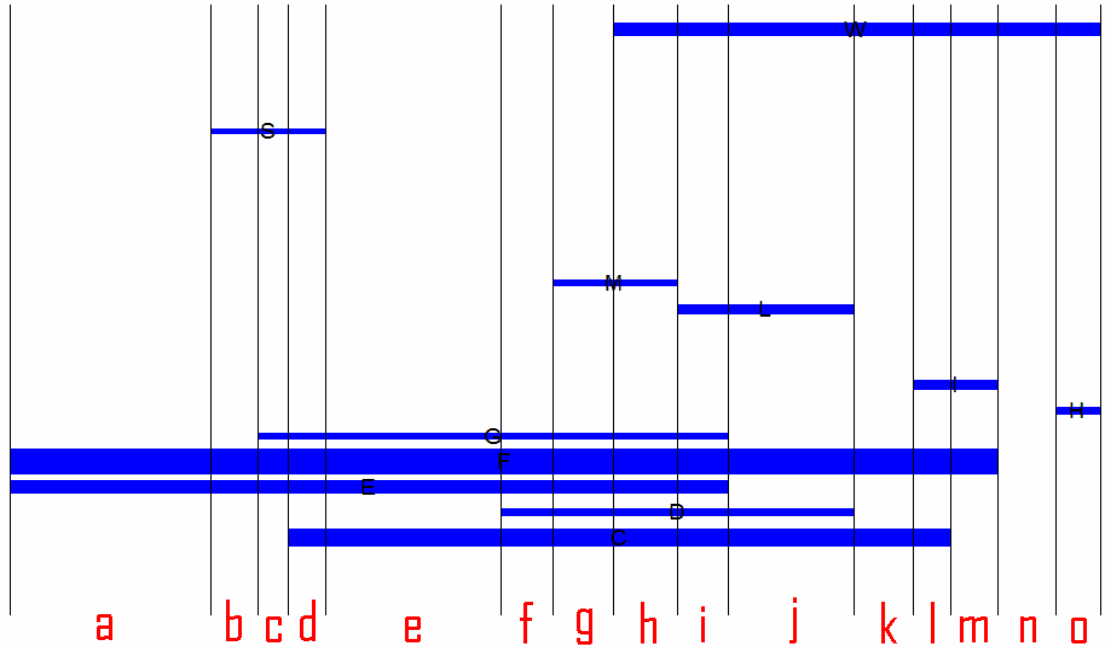


Figure 4.9. Kinetemes for a single actuator in joint angle space.

The number of new kinetemes is at most twice the number of original kinetemes. The new kinetemes are used to represent subparts of the original kinetemes in the motion patterns. An example from Figure 4.9 is the original kineteme **S** that becomes the sequence of subparts **bcd**. This way, every instance of **S** in a motion pattern is replaced by the sequence **bcd**. The attributes of the new kinetemes are retrieved from the original kinetemes they belong to. This way, an original kineteme becomes its sequence of subparts and every instance of an original kineteme symbol in a motion pattern is replaced by its sequence of subpart symbols.

The inference of the CFG that generates motion patterns for actuator *i* in all actions involves the application of sequential learning to a string that is the concatenation of

all these motion patterns: $\langle p_i^{a1}\ p_i^{a2}\ \ldots\ p_i^{ak}\rangle$, where $\langle\rangle$ denotes the concatenation operation and k is the number of actions a such that actuator $i \in I^a$. However, the counting of occurrences of digrams does not consider digrams with symbols at the borders of two different consecutive patterns. For example, a set of two motion patterns **BAC** and **DACD** is concatenated as **BACDACD**, but the first occurrence of digram **CD** is not considered. This way, the ordering of concatenation of motion patterns does not affect the inferred grammar.

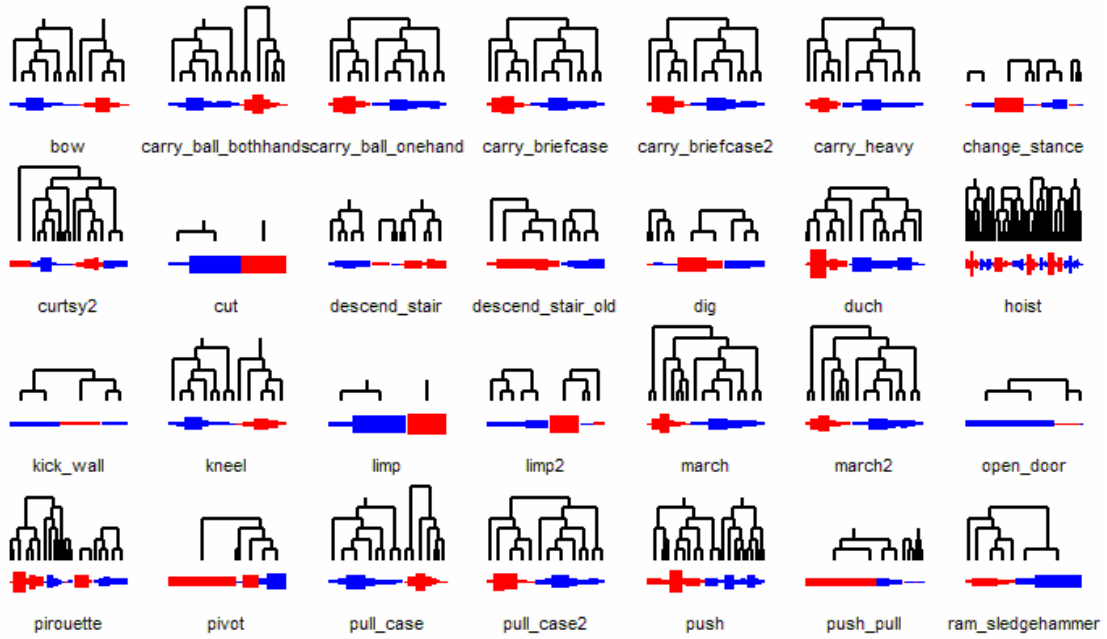


Figure 4.10. Morphological grammar for a single actuator.

The morphological grammar induced for a single actuator is an additional structure that compactly represents all possible motion patterns for this actuator, as Figure 4.10 shows. The grammar is able to generate any movement in the praxicon and to aid the analysis of an unknown (possibly novel) movement. Further, based on the new kinetemes, the grammar explicitly considers the intersections between original kinetemes. This leads to an important aspect of morpho-syntax that is the discovery of

common motions in different actions. The morpho-syntactic process is applied to obtain morphological grammars for each actuator in the articulated human body model (see Appendix C).

Conclusion

In this chapter, we discussed the morphological part of our linguistic framework for the modeling and learning of human activity representations. In this part, we associate each action with a novel formal grammar system. Our heuristic parallel algorithm infers a grammar system without any structural information about the components or component languages.

We presented a human movement representation considering the variability in the set of active joints for different activities. Our representation explicitly contains the set of joints (degrees of freedom) actually responsible for achieving the goal aimed by the activity, the synchronization rules modeling coordination among these actuators, and the motion performed by each participating actuator.

Towards the discovery of a sensory-motor Human Activity Language, we presented the steps required for the construction of a praxicon. A praxicon is the kinematic analogous of a lexicon in spoken language. We learned a large praxicon through the inference of the grammar systems corresponding to a large set of actions. The learned templates of human action allow the mining of strategies of movement. This leads to the syntax of human activity, another part of our linguistic framework, and will have implications in the parsing of human action.

Another important issue concerning the non-arbitrary mapping of motion data to concrete concepts (associated with human action) is the grounding of symbolic

reasoning systems. A logic-based conceptual system is grounded in sensory-motor information through this mapping. Therefore, our linguistic framework is another way to attach meaning to a conceptual reasoning system.

Our framework was able to infer movement patterns that closely model the original movement. The patterns provide high-level and explicit information about the meaning of each human activity. Therefore, our approach was successful in both representational and learning aspects, serving as tool to parse movement, learn patterns, and to generate actions.

Chapter 5: Syntax

The syntax of human activities involves the construction of sentences using action morphemes. A sentence consists of a group of entities. In this sense, a sentence may range from a single action morpheme to a sequence of sets of morphemes. The sets of morphemes represent simultaneous actions and a sequence of movements relates to the causal concatenation of activities This way, our intention is to identify which entities constitute a single morpheme sentence (nuclear syntax) and to study the mechanisms of composing sets of morphemes (parallel syntax) and of connecting these sets into sequences (sequential syntax).

Nuclear Syntax

A single action morpheme sentence is composed of entities that are implicit in any motion. These entities are a central part of an action that we refer as nuclear-syntax. For didactical purposes, we identify these entities as analogs to lexical categories: nouns, adjectives, verbs, and adverbs. An action is represented by a word that has the structure of a sentence: the agent or subject is a set of active body parts (noun), and the action or predicate is the motion of those parts (verb). In many such words, the action is transitive and involves an object or another patient body part.

Nouns and Adjectives

In a sentence, a noun represents the subjects performing an activity or objects receiving an activity. A noun in a single action sentence corresponds to the essential body parts active during the execution of a human activity and to the possible objects involved passively in the action (including patient body parts). The body parts are

equivalent to actuators of the articulated body model. Therefore, a noun (active body parts) is retrieved from the set of essential actuators in the action morpheme. This set may be represented as a binary string with the same size of the set of all actuators. Each element of this string encodes the inclusion of a particular joint actuator in this set. Given the morphology of each action in our database, we may find a matrix where each column is a binary string encoding the noun for a different action, as shown in Figure 5.1. This way, the rows of this matrix correspond to actuators. The *noun matrix* is a low-level structure containing the vocabulary of nouns for a praxicon.

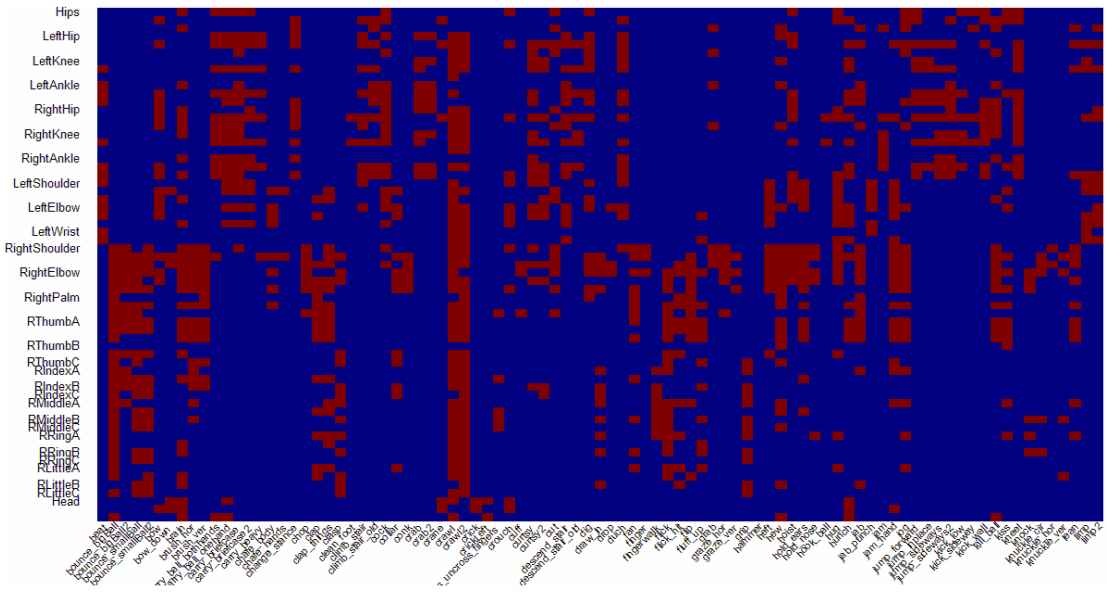


Figure 5.1. Matrix with nouns for a praxicon.

From the morphemes of our motion database, we have extracted a set of about 200 binary strings representing the HAL nouns in the most basic level for each action. Using the noun matrix of a praxicon, we can infer a grammar that resembles the topology of the body model. High-level nouns correspond to high-level non-terminals in this grammar associated with body parts such as lower limb, hand, and head.

The initial posture of an action is analogous to an adjective which further describes (modifies) the active body parts (nouns) in the sentence. The initial pose of an action is retrieved from a morpheme as the initial joint angle of the first kineteme in the motion pattern of each essential actuator. Higher-level adjectives represent usual initial postures such as sit, stand, and lie.

Verbs and Adverbs

A motion verb represents the changes each active actuator experiences during the action execution. The human activity verbs are obtained from the motion patterns in the action morphemes.

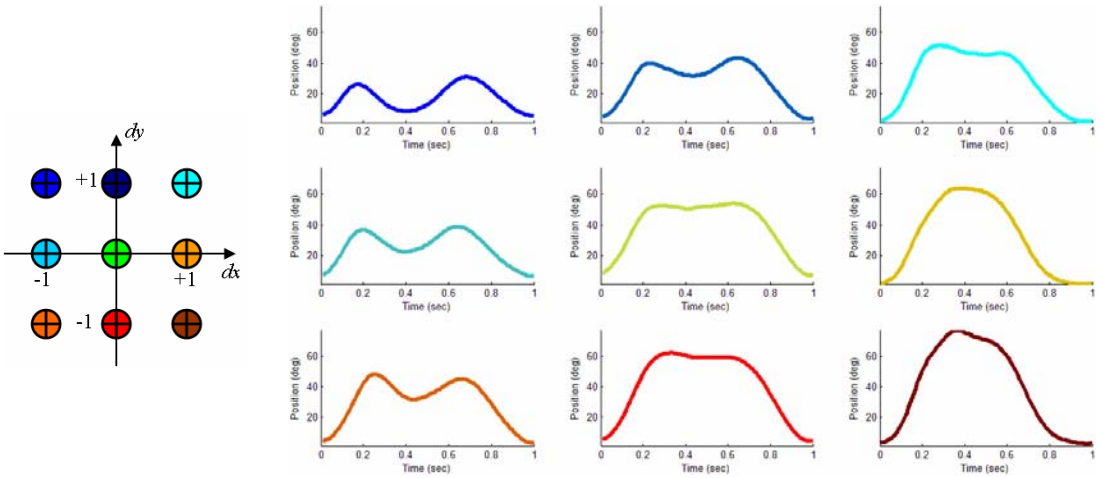
An adverb models the variation in the execution of each motion segment in a verb. The adverb modifies the verb with the purpose of generalizing the motion. For example, one instance of a "reach with your hand" action corresponds to a morpheme that models the movement required to touch something at a specific location. To generalize this action to any location, the motion of a segment is represented in a space with a reduced dimensionality. Each dimension in this reduced space represents a parameter such as location, speed, and force that models the variability of an action. The usual dimensionality reduction methods (e.g., PCA, ICA) learn meaningless parameters (semantic problem). Further, the parameters inferred for one action have no relationship to the parameters learnt for another action (universality problem). The semantic problem is addressed in our approach to adverbs by explicitly selecting a meaningful parametric space "a priori". We suggest an intuitive set of parameters that consist of origin (o_x, o_y, o_z), destination (d_x, d_y, d_z), speed (s), and a resistance force (f). For example, the "reach with your hand" action is specified by the start position

of the hand (possibly at a resting location), the end position of the hand (where the object is located at), how fast the movement is performed, and any resistance force involved in the action. A resistance force could be the weight of an object being carried or the slope in an inclined walk.

Formally, an adverb is an eight-element vector [o_x, o_y, o_z, d_x, d_y, d_z, s, f] used to further describe any action. This way, the universality problem is solved by using the same parametric space to represent adverbs of different actions. The adverbial modeling process consists in the mapping from parametric space to motion. Consequently, this process provides an immediate mapping between extrinsic coordinates (Cartesian 3D locations) and intrinsic coordinates (joint angles). This way, the user of a data-driven computer animation system specifies a location, even in terms of virtual objects, and the motion adverb is computed appropriately.

The modeling process involves the interpolation of sample motions in the parametric space. Although we have studied this process in several actions (e.g., reach, sit, kick, walk, inclined walk, run) and for many parameters with similar results, we only discuss here the experiments about the kick and walk actions. The kick action was analyzed according to location parameters (d_x, d_y), while the walk forward action was investigated for the speed parameter (s). The input required by this process consists of motion samples distributed in the parametric space. For the kick action, we captured kick motions with the right leg from a single resting stance position to several target destinations placed on different horizontal (d_x) and vertical (d_y) locations, as shown in Figure 5.2. The horizontal location varies as left (d_x = -1), center (d_x = 0), and right (d_x = +1). The vertical location varies as bottom (d_y = -1), center (d_y = 0), and top

(d_y = +1). The remaining parameters are constant and the targets were located in a vertical plane.



(a) Target locations.

(b) The "right knee flexion-extension".

Figure 5.2. The kick action for distributed parameters.

The interpolation of sample motions is performed for each actuator and each time instant independently. In other words, for a specific actuator i, we consider only a single time instant t to compute an interpolation model for other parametric points at the same instant t. This way, we denote the motion $J_i(t)$ with parameters (d_x, d_y) as $M_i^t(d_x, d_y)$. We discuss a quadratic model $M_i^t(d_x, d_y) = Ad_x^2 + Bd_y^2 + Cd_xd_y + Dd_x + Ed_y + F$ for the interpolation. Our goal is to find the model components A, B, C, D, E, and F for every time t and for each single actuator i. This way, we need at least 6 equations to fully determine these variables and, consequently, the minimum of 6 motion samples is required. From the given motion samples in our experiment, we have 9 equations. This way, the model components are computed by a least squares method that solves an over-determined linear system. The model components for every time t are shown in Figure 5.3. Note that component F is motion $M_i^t(0, 0)$.

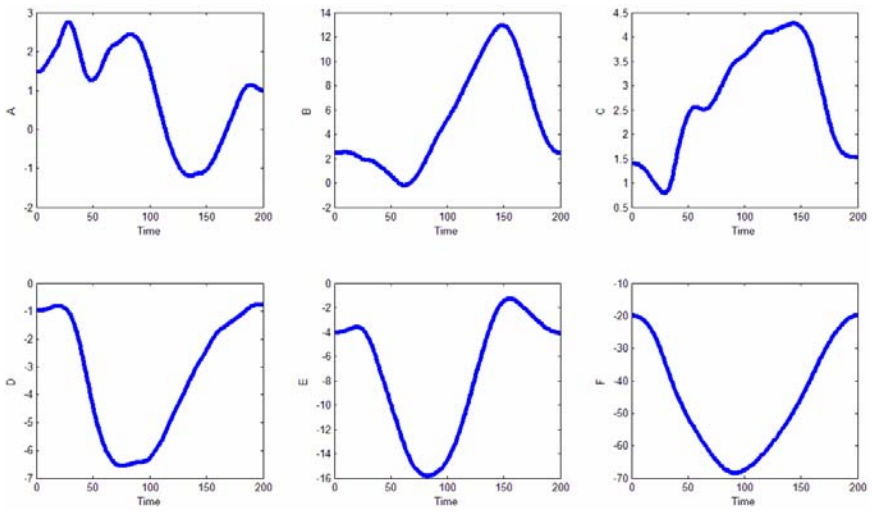


Figure 5.3. Quadratic components for generalization of motion in the "right hip flexion-extension" actuator in the kick action.

Once the model components are obtained, we may compute motions for any point in the adverbial parametric space. Figure 5.4 shows regularly spaced interpolated motions. The colored curves represent original sample motions superimposed over the interpolated motions. The average error (absolute difference) of the interpolated reconstruction for all samples was 1.3°.

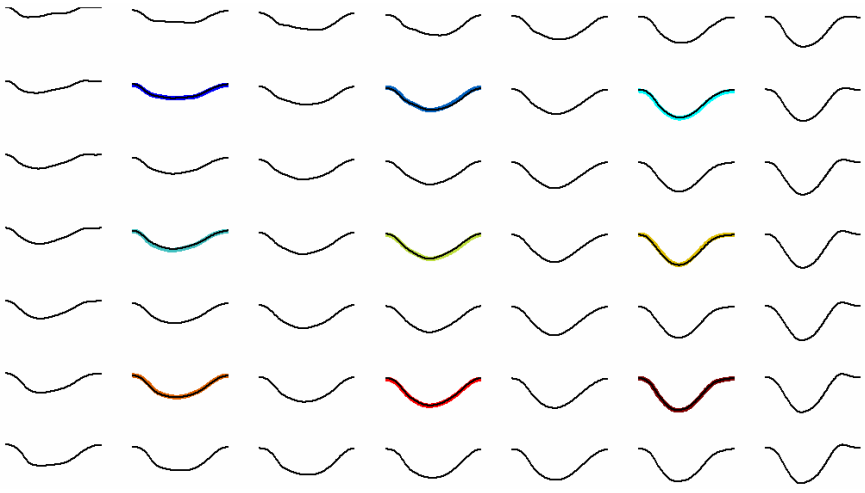


Figure 5.4. Interpolated motions using a quadratic model.

We discussed above an modeling process for location parameters $\{o_x, o_y, o_z, d_x, d_y, d_z\}$ of the adverbial space. For the speed parameter $\{s\}$, we suggest a different approach. To investigate this parameter, we use the walk forward action. In this case, our sample motions are walk actions at 40 different speeds regularly spaced from 0.1mph to 4.0mph. Figure 5.5 shows these sample motions for the "right knee flexion-extension" actuator. The motion is presented as angular position, angular velocity, angular acceleration, and angular jerk. Actually, we show a single cycle of the walk action normalized in time. The motion curves are colored according to the speed such that colder colors (e.g., blue) display slower walks and warmer colors (e.g., red) display faster walks.

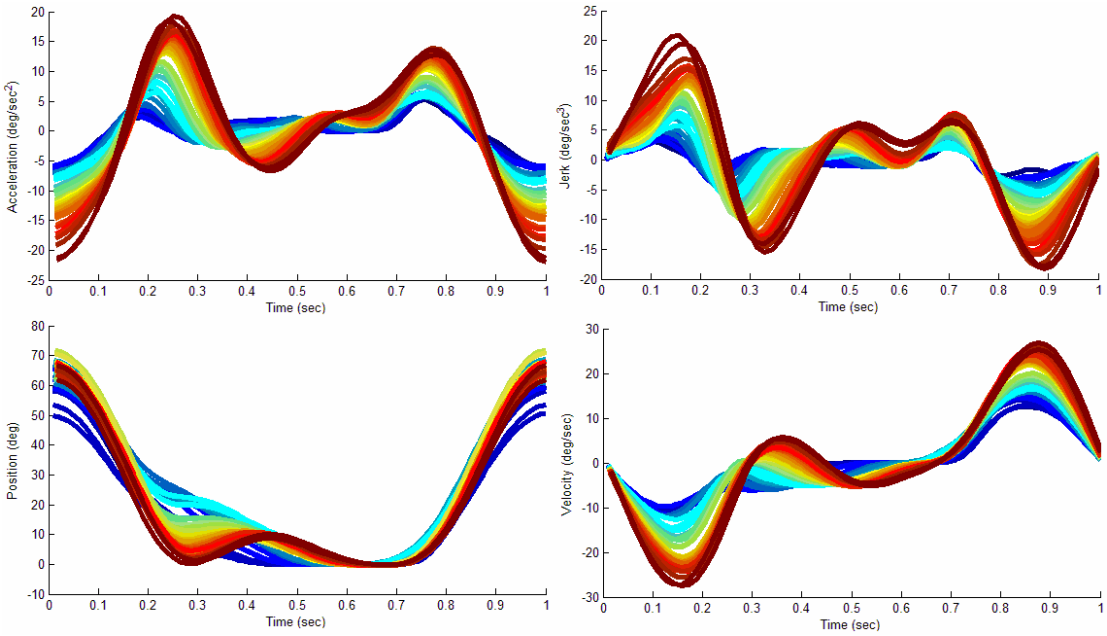


Figure 5.5. Walk action at different speeds.

Note that, for the speed varying motions, the most variability in the curves is at the maxima and minima points. Coincidently and fortunately, these points are the borders of motion segments. With this in mind, we aim at modeling how these extreme points

behave in time and space (i.e., position, velocity, acceleration, and jerk) according to the speed parameter. Let's consider a single extreme point e and denote its time and space for a walk motion at speed s as $t_e(s)$ and $q_e(s)$, respectively. We discovered from the motion data that these functions are fairly well modeled by a line, as Figure 5.6 shows. This way, only two values are required to represent each extreme point behavior according to the speed parameter.

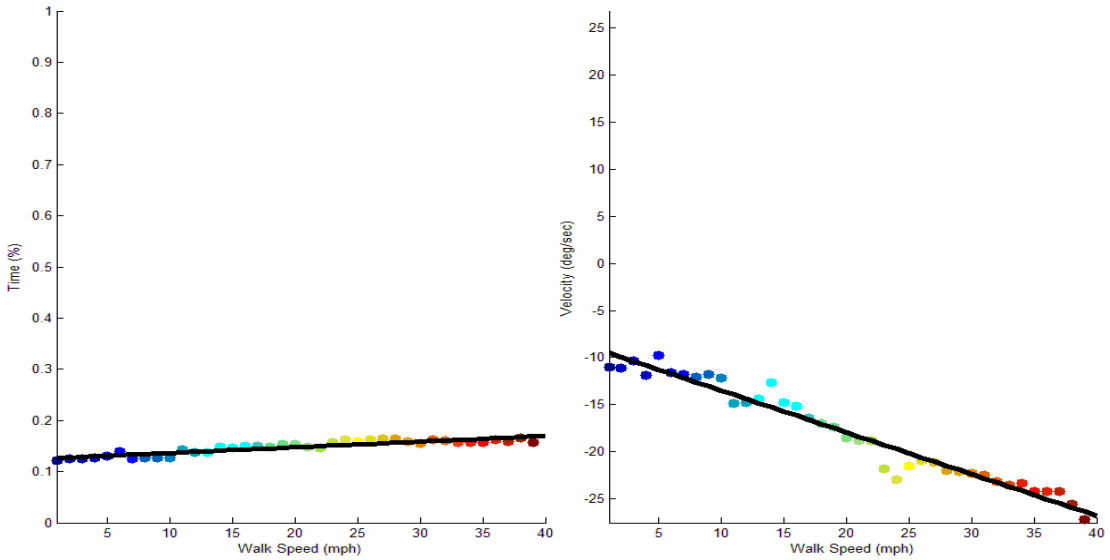


Figure 5.6. Time and space functions of an extreme point at varying speeds of the walk action.

This modeling process was evaluated for several actuators and various extreme points. For each extreme point e, the evaluation consisted in using a subset S of sample speeds to compute the lines l_e^t and l_e^q modeling $t_e(s)$ and $q_e(s)$ for all $s \in S$, respectively. Once these lines are computed, the average values $|l_e^t(s) - t_e(s)|$ and $|l_e^q(s) - q_e(s)|$ for all sample speeds in S are the respective model error. Figure 5.7 shows the model error of time (left) and space (right) for one extreme point. Each

graph shows the model error for a decreasing number of sample speeds. For a reasonable amount of sample speeds, the spatial model error is less than 1°.

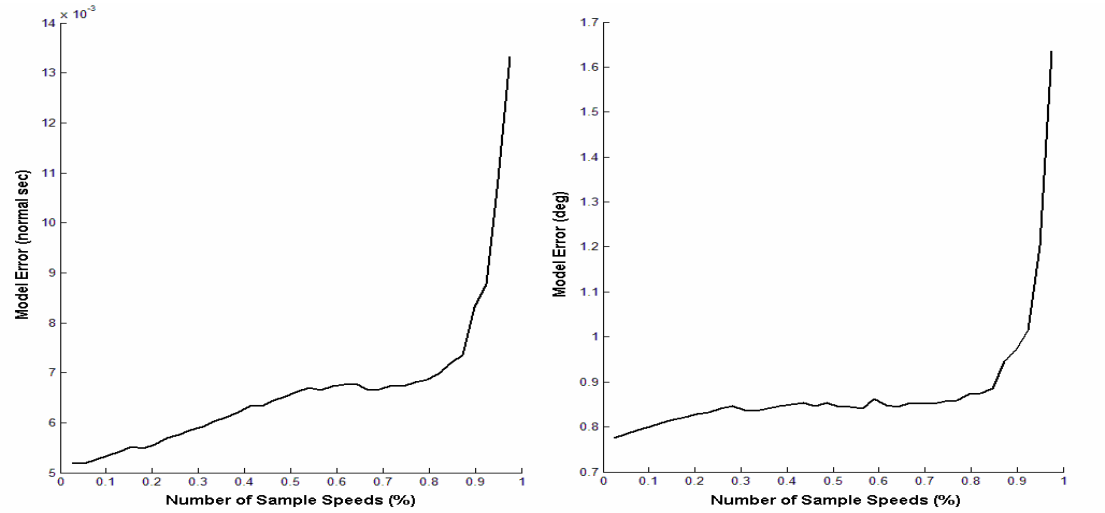


Figure 5.7. Model error increases with less sample speeds.

Spatio-Temporal Syntax

The lexical categories proposed for HAL compose a nuclear syntax. A HAL sentence S → NP VP consists of noun phrase (noun + adjective) and verbal phrase (verb + adverb), where NP → N Adj and VP → V Adv, as shown in Figure 5.8. However, the organization of human movement is also simultaneous and sequential. This way, the nuclear syntax expands to parallel and sequential syntax.

The parallel syntax concerns activities performed simultaneously represented by parallel sentences $S_{t,\,j}$ and $S_{t,\,j+1}$. This syntax constrains the respective nouns of the parallel sentences to be different: $N_{t,j} \neq N_{t,j+1}$. This constraint states that simultaneous movement must be performed by different body parts. For example, a person may walk and wave at concurrently. However, one cannot whistle and chew gum at the same time!

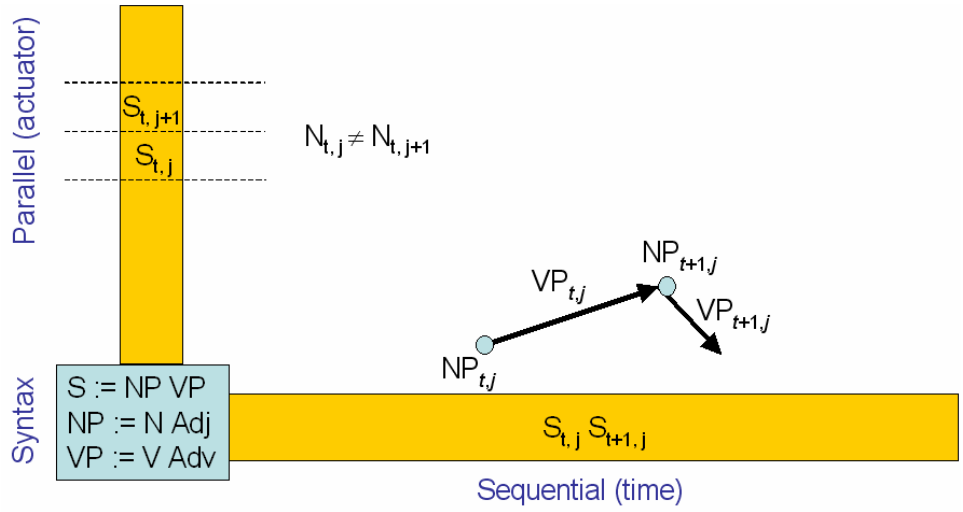


Figure 5.8. Nuclear, parallel, and sequential syntax.

The temporal sequential combination of action sentences ($S_{t,j}\ S_{t+1,j}$) must obey the cause and effect rule. The HAL noun phrase must experience the verb cause and the joint configuration effect must lead to a posture corresponding to the noun phrase of the next sentence. Considering noun phrases as points and verb phrases as vectors in the same space, the cause and effect rule becomes $NP_{t,j} + VP_{t,j} = NP_{t+1,j}$. The cause and effect rule is physically consistent and embeds the ordering concept of syntax.

Parallel Syntax

Parallel syntax addresses the possible ways to combine different action morphemes into a set of morphemes that could be performed simultaneously. Basically, the main constraint imposed by parallel syntax involves the essential actuators. To merge two action morphemes for actions a_1 and a_2 into a parallel set of morphemes, their sets of essential actuators I^{a1} and I^{a2} need to have an empty intersection. In other words, the two action morphemes cannot share any essential actuator. This rule may be implemented as a constraint matrix C. For each pair of actions a_1 and a_2 in a praxicon, if $I^{a1} \cap I^{a2} = \varnothing$, the matrix entry $C(a_1, a_2)$ is true; otherwise, the matrix entry is false.

The constraint matrix explicitly stores which pairs of morphemes could be merged as simultaneous activities, shown as red entries in Figure 5.9. More sophisticated inferences could also be performed using this structure. For example, transforming this matrix into a graph, cliques correspond to groups of action morphemes that may be executed at the same time.

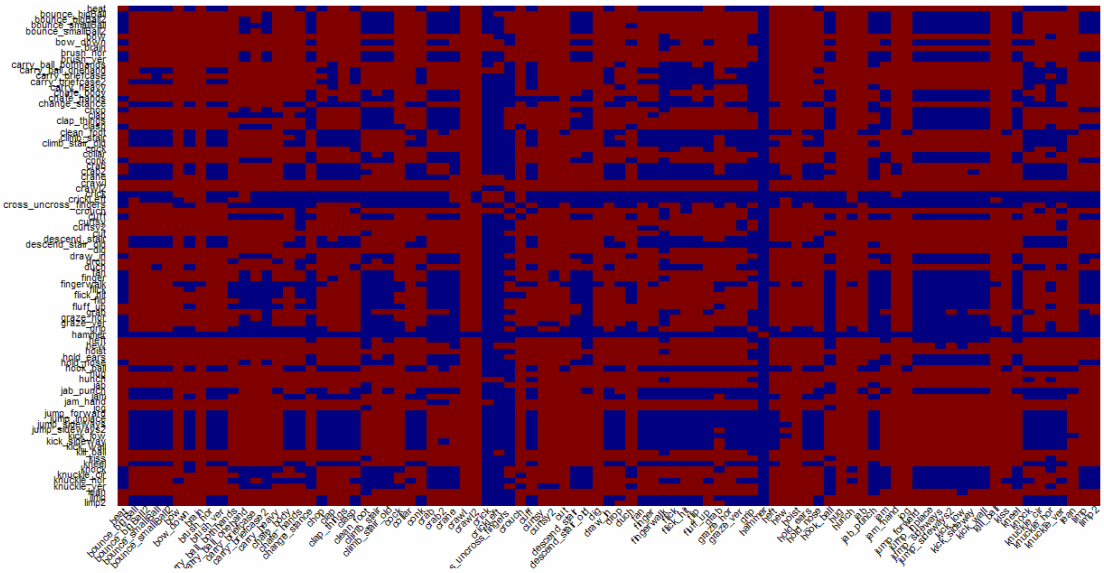


Figure 5.9. A constraint matrix for simultaneous actions.

Sequential Syntax

In speech, the temporal organization is a pre-syntax since this neural preplanning of motor action is what syntax uses to execute an utterance. Actions of the physical body provide a metaphor for the hierarchical structure of language. The precise muscle timing (pre-syntax) makes it possible to produce countless actions that differ in great or small ways. The lexical units are arranged into sequences to form sentences. A sentence is a sequence of actions that achieve some purpose.

The cause and effect rule is physically consistent and embeds the ordering concept of syntax. The body pose must experience the motion cause and the effect leads to a posture in the next sentence. Sequential syntax concerns the concatenation of actions or, more formally, the connection of sets of action morphemes (from parallel syntax) to form sequences of movement.

Consider a single actuator i, if i belongs to the sets I^{a1} and I^{a2} of essential actuators of two action morphemes a_1 and a_2, respectively, the sequential concatenation of these two morphemes is only feasible if there is a transition from one motion pattern p_{a1} to the other p_{a2}. Such a transition may be obtained from the morphological grammar G_i of actuator i (as discussed in morpho-syntax). Any non-terminals or terminals in G_i shared by both motion patterns p_{a1} and p_{a2} give rise to a possible transition. Consequently, the two morphemes a_1 and a_2 have a feasible concatenation with respect to actuator i. This way, two sets of action morphemes may be sequentially connected only if they have a feasible concatenation with respect to all actuators contained in the intersection of their sets of essential actuators. Figure 5.10 displays the motion patterns of two action morphemes and their respective morphological grammar entries. The two patterns share kinetemes and, consequently, a transition exists between the two morphemes.

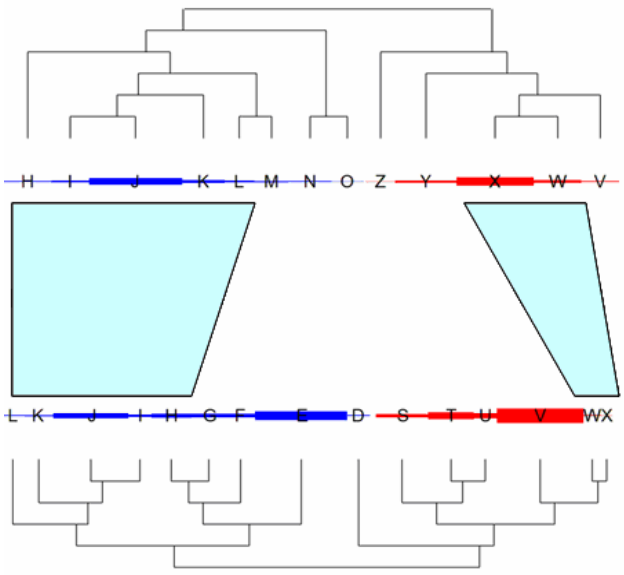


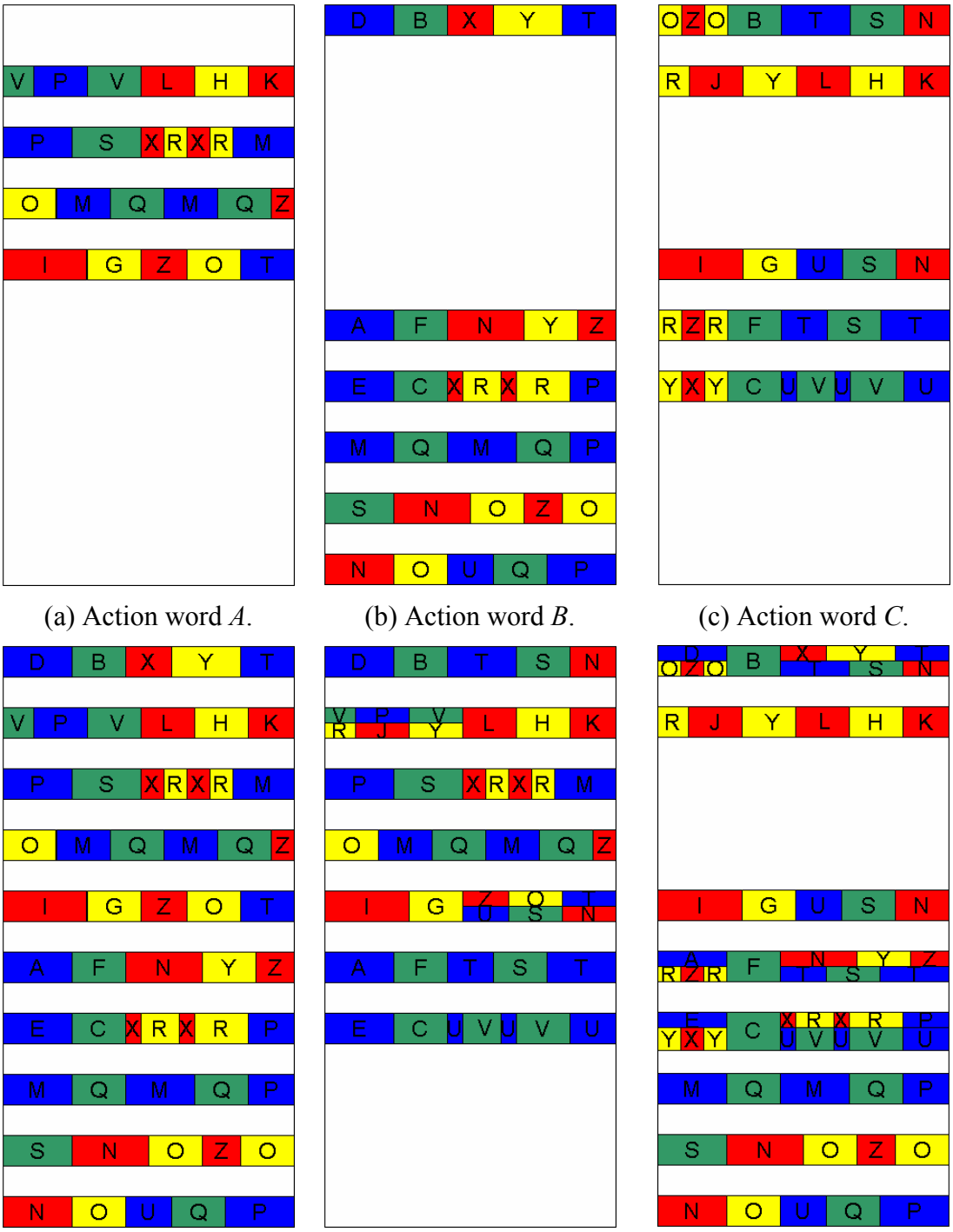
Figure 5.10. Possible transitions between two morphemes.

The lexical units are arranged into sequences to form sentences. A sentence is a sequence of actions that achieve some purpose. In written language, sentences are delimited by punctuation. Analogously, the action language delimits sentences using motionless actions. In general, a conjunctive action is performed between two actions, where a conjunctive action is any preparatory movement that leads to an initial position required by the next sentence.

Conclusion

In this chapter, we discussed the sentence formation process. In Figure 5.11, we illustrate this process using three action words *A*, *B*, and *C* (see Figure 5.11a-c). Since *A* and *B* have disjoint sets of essential actuators, we can form the simultaneous sentence *A* || *B* (see Figure 5.11d). Action words *A* and *C* share two essential actuators and, consequently, only sequential composition applies. Although there are transitions

for both actuators, they are not concurrent and the sequential sentence $A \bullet C$ is not feasible (see Figure 5.11e). The sequential sentence $B \bullet C$ is shown in Figure 5.11f.



(a) Action word *A*. (b) Action word *B*. (c) Action word *C*.

(d) Sentence $A \parallel B$. (e) Sentence $A \bullet C$ is not feasible. (f) Sentence $B \bullet C$.

Figure 5.11. Sentence formation process.

Chapter 6: Conclusion

Perhaps, a more general cognitive capacity allows us to decode highly coded signals. A reasonable hypothesis would be that there is little difference between the visual and the speech realms in this regard. The visual stimuli available to the brain do not offer a stable code of information. The brain extracts the constant invariant features of objects from the perpetually changing flood of information it receives from them. Further, what is being perceived and apprehended is the message itself, not static target end states of the articulators.

In the scientific discussions about perception and reasoning in cognitive systems, a debate on "signals vs. symbols" has been going on for quite some time. Specifically, where do signals end and where do symbols begin? In other words, what are the boundaries among signal processing, computer vision (and audition), and artificial intelligence? Our work has demonstrated that this debate may not be very fruitful, because signals and symbols acquire their meaning depending on the operations that we apply to them.

Starting from motion capture measurements (signals), we extract symbols as early as possible and then utilize a symbolic framework to learn a first language of human movement. Following the framework of modern linguistics [Jackendoff, 1997], we studied the kinetology, morphology and syntax of this new language (and did not mention at all semantics and pragmatics). We discussed the nouns, adjectives, verbs, and adverbs of this language. Our work on adverbial modeling brings forward additional parameters, such as location, speed, and force. Like modern day archaeologists working from a papyrus containing a series of actiongrams, we

decipher the underlying language, using computer science techniques, by discovering the structure in each joint and among joints.

We must emphasize that we have merely scratched the top of an iceberg. We simply demonstrated that there exists a language of human activity by empirically constructing one such language out of large amounts of data. Our kinetology was among the simplest possible, yet rich enough to provide an interesting structure. It should be clear that there is a trade-off between the complexity of the kinetology and the complexity of the grammar. Very simple kinetemes give rise to complex grammars, while more structured kinetemes produce simpler grammars. A recent effort is to develop a spectral kinetology, where the kinetemes are basic functions (wavelets) linked with a number of parameters for each joint. The idea is that a single wavelet in conjunction with the provided parameters will produce the whole function (movement) of a synergy of joints. This approach will give rise to simpler grammars.

Applications of HAL in various areas are bringing a novel viewpoint. Given HAL, the problem of visual surveillance and video analysis becomes one of translation from image representations to HAL, i.e., action understanding involves motor representations (at a higher or abstract level that the language provides). Most importantly, HAL addresses a fundamental research issue in cognitive science and artificial intelligence, that of the mechanisms for the combination of sensory-motor information and concepts for understanding others and for communicating one's own intentions. HAL suggests an empirical approach for discovering the "languages" of action and vision and the correspondences with natural language, by collecting body movement measurements, visual object information, and associated linguistic

descriptions from interacting human subjects. Using such data, one can imagine that in the near future the community will move towards the creation of the praxicon, a computational resource that associates the lexicon (words/concepts) with corresponding motoric and sensory representations and that is enriched with information on co-occurrence patterns among the concepts for forming higher-level complex concepts. This praxicon will bring us closer to understanding human thought while significantly enhancing software that acquires "meaning". It also suggests the new idea of achieving artificial intelligence by measuring (structuring, parsing, and analyzing) human behavior.

From a methodological viewpoint, this dissertation introduced a new way of achieving an artificial cognitive system through the study of human action, or to be more precise, through the study of the sensory-motor system. We believe this study represented initial steps of one approach towards conceptual grounding. The closure of this semantic gap will lead to the foundation of concepts into a non-arbitrary meaningful symbolic representation based on sensory-motor intelligence. This representation will serve to the interests of reasoning in higher-level tasks and open the way to more effective techniques with powerful applications.

Humans have been studying the spoken and written languages for thousands of years. It is not clear how long it will take to map out the murky depths of a Human Activity Language. We hope that HAL is a step in the right direction.

Appendix A: Concrete Verbs

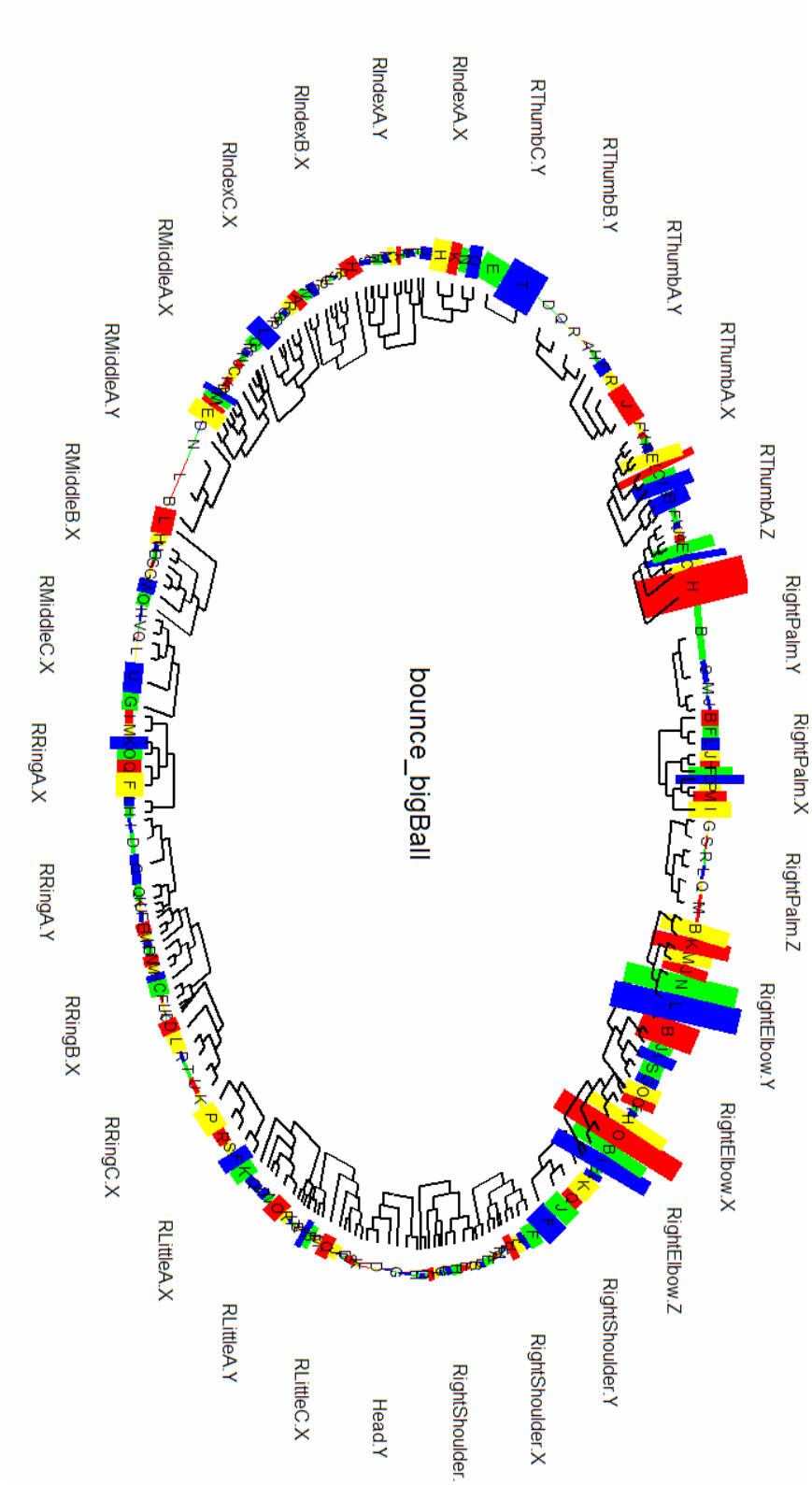
WordNet is a lexical database for the English language that organizes words into a semantic network [Fellbaum, 1998]. A semantic network links words according to relations such as synonym-antonym, hyponym-hypernym (kind of), and meronym-holonym (part of). We used the relations between verbs in the WordNet semantic network to categorize English verbs associated with observable voluntary actions as *concrete verbs*. Our praxicon was based on the subset of concrete verbs enumerated below. Each synset (set of synonyms) is presented in the following form: "{ verb words } = gloss;":

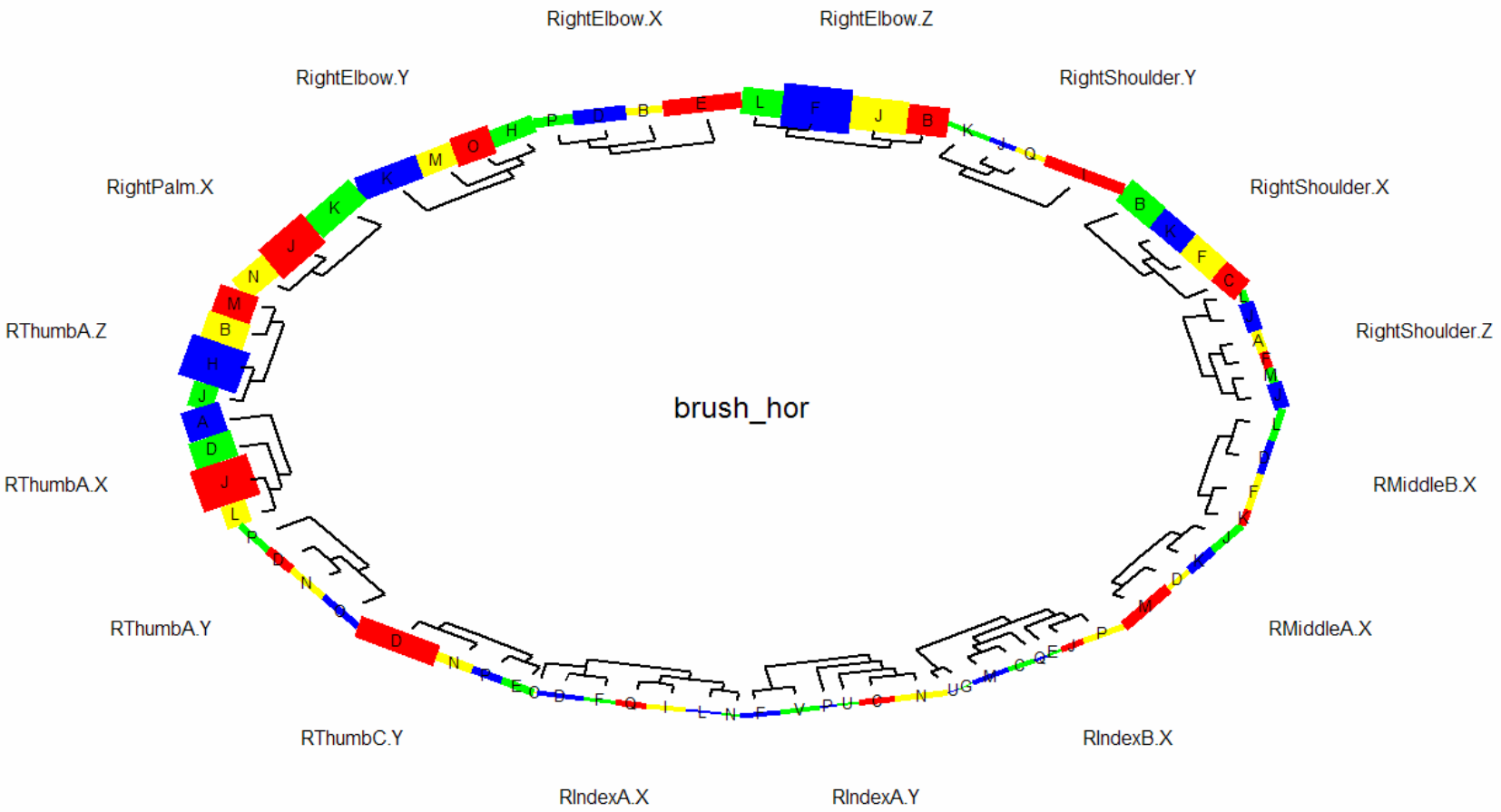
{ beat } = hit repeatedly;
{ beckon wave } = signal with the hands or nod;
{ bounce } = hit something so that it bounces;
{ bow } = bend the head or the upper part of the body in a gesture of respect or greeting;
{ bow bow_down } = bend one's knee or body, or lower one's head;
{ brain } = hit on the head;
{ brandish flourish wave } = move or swing back and forth;
{ brush } = touch lightly and briefly;
{ carry } = bear or be able to bear the weight, pressure, or responsibility of;
{ catch grab take_hold_of } = take hold of so as to seize or restrain or stop the motion of;
{ chafe } = warm by rubbing, as with the hands;
{ chop } = strike sharply, as in some sports;
{ clap } = strike with the flat of the hand; usually in a friendly way, as in encouragement or greeting;
{ clap } = strike together so as to produce a sharp percussive noise;
{ clasp } = grasp firmly;
{ cock } = tilt or slant to one side;
{ collar } = seize by the neck or collar;
{ conk } = hit, especially on the head;
{ crab } = scurry sideways like a crab;
{ crane stretch_out } = stretch (the neck) so as to see better;
{ crawl creep } = move slowly; in the case of people or animals with the body near the ground;
{ crick } = twist the head into a strained position;
{ cross } = fold so as to resemble a cross;
{ crouch stoop bend bow } = bend one's back forward from the waist on down;
{ cuff whomp } = hit with the hand;
{ curtsy bob } = make a curtsy; usually done only by girls and women; as a sign of respect;
{ curtsy curtsey } = a gesture of respectful greeting, for women;
{ cut } = move (one's fist);
{ draw_in retract } = pull inward or towards a center;
{ drop } = let fall to the ground;
{ duck } = to move (the head or body) quickly downwards or away;

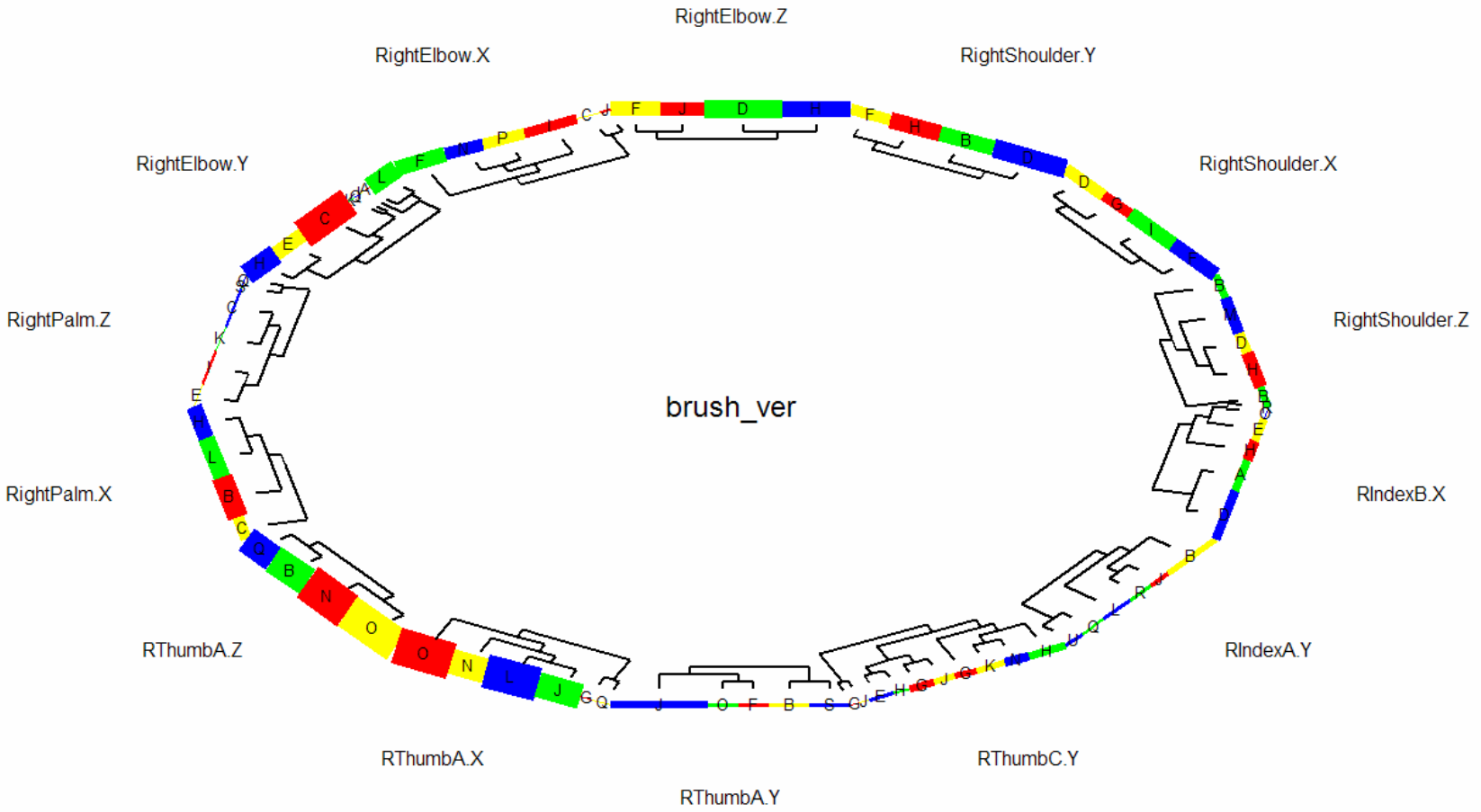
{ embrace hug bosom squeeze } = hug, usually with fondness;
{ fan } = agitate the air;
{ finger thumb } = feel or handle with the fingers;
{ flick } = touch or hit with a light, quick blow;
{ flip twitch } = toss with a sharp movement so as to cause to turn over in the air;
{ fluff_up plump_up shake_up } = make fuller by shaking;
{ grab } = make a grasping or snatching motion with the hand;
{ graze crease rake } = scrape gently;
{ grip } = hold fast or firmly;
{ heft } = test the weight of something by lifting it;
{ hew } = strike with an axe; cut down, strike;
{ hoist } = move from one place to another by lifting;
{ hold } = cover as for protection against noise or smell;
{ hold take_hold } = have or hold in one's hands or grip;
{ hook } = secure with the foot;
{ hunch hump hunch_forward hunch_over } = arch one's back;
{ jab } = strike or punch quick and short blows;
{ jam } = push down forcibly;
{ jog } = run for exercise;
{ jump leap bound spring } = move forward by leaps and bounds;
{ kick } = strike with the foot;
{ kick } = drive or propel with the foot;
{ kill } = hit with great force;
{ kiss buss osculate } = touch with the lips or press the lips (against someone's mouth or other body part) as an expression of love, greeting, etc;
{ kneel } = rest one's weight on one's knees;
{ knock } = rap with the knuckles;
{ knuckle } = press or rub with the knuckles;
{ limp hobble hitch } = walk impeded by some physical limitation or injury;
{ look_back look_backward } = look towards one's back;
{ lower take_down let_down get_down bring_down } = move something or somebody to a lower position;
{ march } = walk fast, with regular or measured steps; walk with a stride;
{ nod } = lower and raise the head, as to indicate assent or agreement or confirmation;
{ nuzzle nose } = rub noses;
{ oscillate vibrate } = move or swing from side to side regularly;
{ pace step } = measure (distances) by pacing;
{ palpate feel } = examine (a body part) by palpation;
{ pet } = stroke or caress gently;
{ pick_at pluck_at pull_at } = pluck or pull at with the fingers;
{ pick_up lift_up gather_up } = take and lift upward;
{ pirouette } = do a pirouette, usually as part of a dance;
{ pivot swivel } = turn on a pivot;
{ pluck tweak pull_off pick_off } = pull or pull out sharply;
{ press } = exert pressure or force to or upon;
{ pull } = apply force so as to cause motion towards the source of the motion;
{ pull } = cause to move in a certain direction by exerting a force upon, either physically or in an abstract sense;
{ pull_back } = move to a rearward position; pull towards the back;
{ punch plug } = deliver a quick blow to;
{ push } = press against forcefully without being able to move;
{ push force } = move with force;
{ push_up } = push upward;
{ raise lift elevate get_up bring_up } = raise from a lower to a higher position;
{ ram ram_down pound } = strike or drive against with a heavy impact;
{ roll revolve } = cause to move by turning over or in a circular manner of as if on an axis;

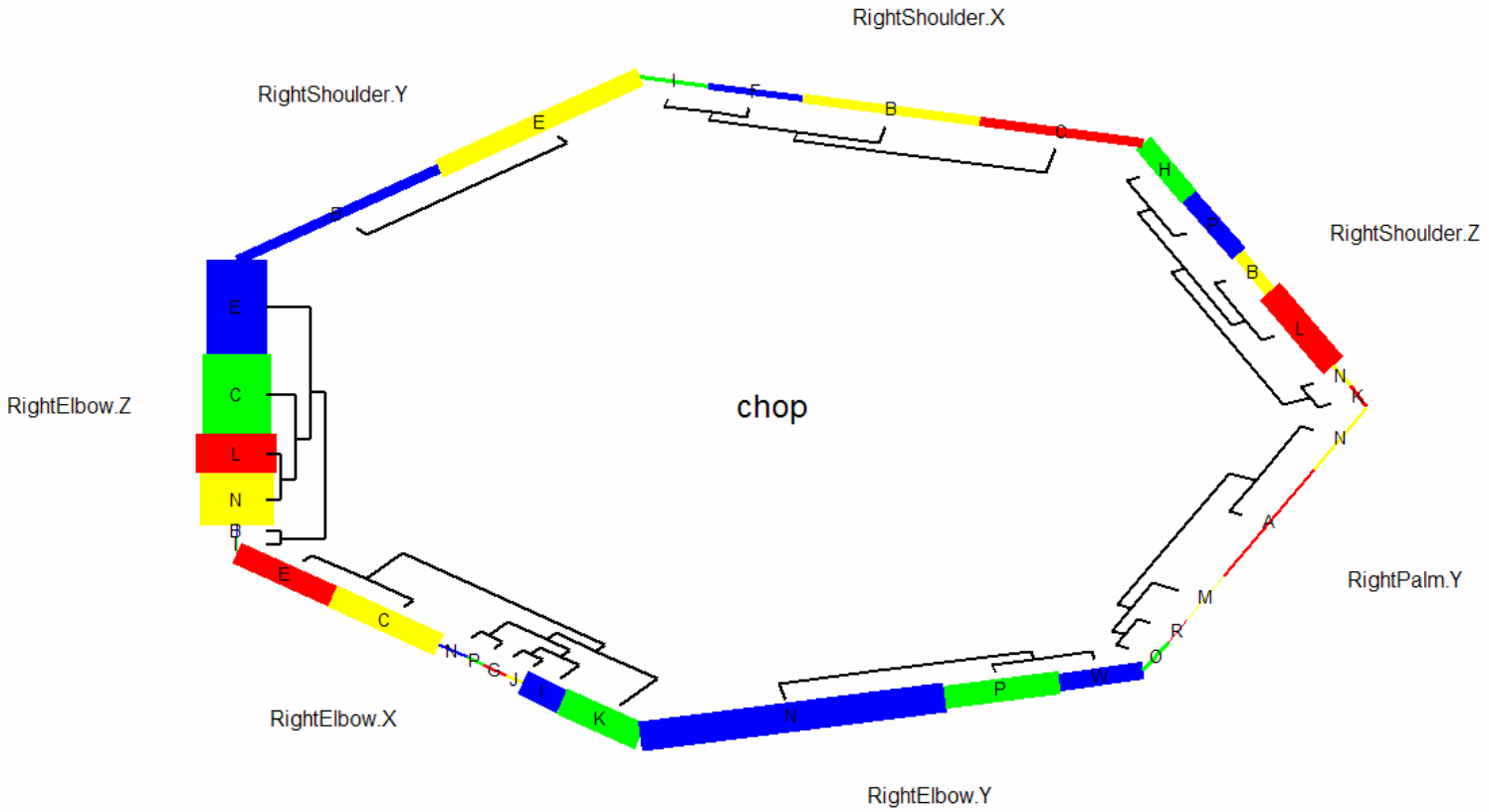
{ rub } = move over something with pressure;
{ run } = move fast by using one's feet, with one foot off the ground at any given time;
{ salute } = greet in a friendly way;
{ salute present } = recognize with a gesture prescribed by a military regulation; assume a prescribed position;
{ scrape kowtow genuflect } = bend the knees and bow in a servile manner;
{ screw drive_in } = cause to penetrate, as with a circular motion;
{ scuff } = poke at with the foot or toe;
{ shake } = shake (a body part) to communicate a greeting, feeling, or cognitive state;
{ shoulder } = lift onto one's shoulders;
{ shuffle scuffle shamble } = walk by dragging one's feet;
{ skim skip skitter } = cause to skip over a surface;
{ slam bang } = strike violently;
{ slap } = hit with something flat, like a paddle or the open hand;
{ slide } = move smoothly along a surface;
{ sling catapult } = hurl as if with a sling;
{ slug slog swig } = strike heavily, especially with the fist or a bat;
{ slump slouch } = assume a drooping posture or carriage;
{ smash } = collide or strike violently and suddenly;
{ sprawl } = sit or lie with one's limbs spread out;
{ spread-eagle } = stand with arms and legs spread out;
{ sprint } = run very fast, usually for a short distance;
{ squash crush squelch mash squeeze } = to compress with violence, out of natural shape or condition;
{ squat crouch scrunch scrunch_up hunker hunker_down } = sit on one's heels;
{ stab jab } = stab or pierce;
{ stand_still } = remain in place; hold still; remain fixed or immobile;
{ step } = move with one's feet in a specific manner;
{ stoop } = carry oneself, often habitually, with head, shoulders, and upper back bent forward;
{ straighten } = get up from a sitting or slouching position;
{ stride } = walk with long steps;
{ stroke fondle } = touch lightly and with affection, with brushing motions;
{ swagger ruffle prance strut sashay cock } = to walk with a lofty proud gait, often in an attempt to impress others;
{ swing } = move in a curve or arc, usually with the intent of hitting;
{ swing sweep swing_out } = make a big sweeping gesture or movement;
{ throw } = project through the air;
{ tip } = cause to tilt;
{ tiptoe tip tippytoe } = walk on one's toes;
{ toe } = walk so that the toes assume an indicated position or direction;
{ tread trample } = tread or stomp heavily or roughly;
{ trot jog clip } = run at a moderately swift pace;
{ tug } = move by pulling hard;
{ turn } = cause to move around or rotate;
{ turn } = cause to move along an axis or into a new direction;
{ turn } = change orientation or direction, also in the abstract sense;
{ turn turn_over } = cause to move around a center so as to show another side of;
{ twitch } = move or pull with a sudden motion;
{ uncross } = change from a crossed to an uncrossed position;
{ volley } = hit before it touches the ground;
{ walk } = use one's feet to advance; advance by steps;
{ whack wham whop wallop } = hit hard;
{ whip lash } = strike as if by whipping;
{ zigzag crank } = travel along a zigzag path;

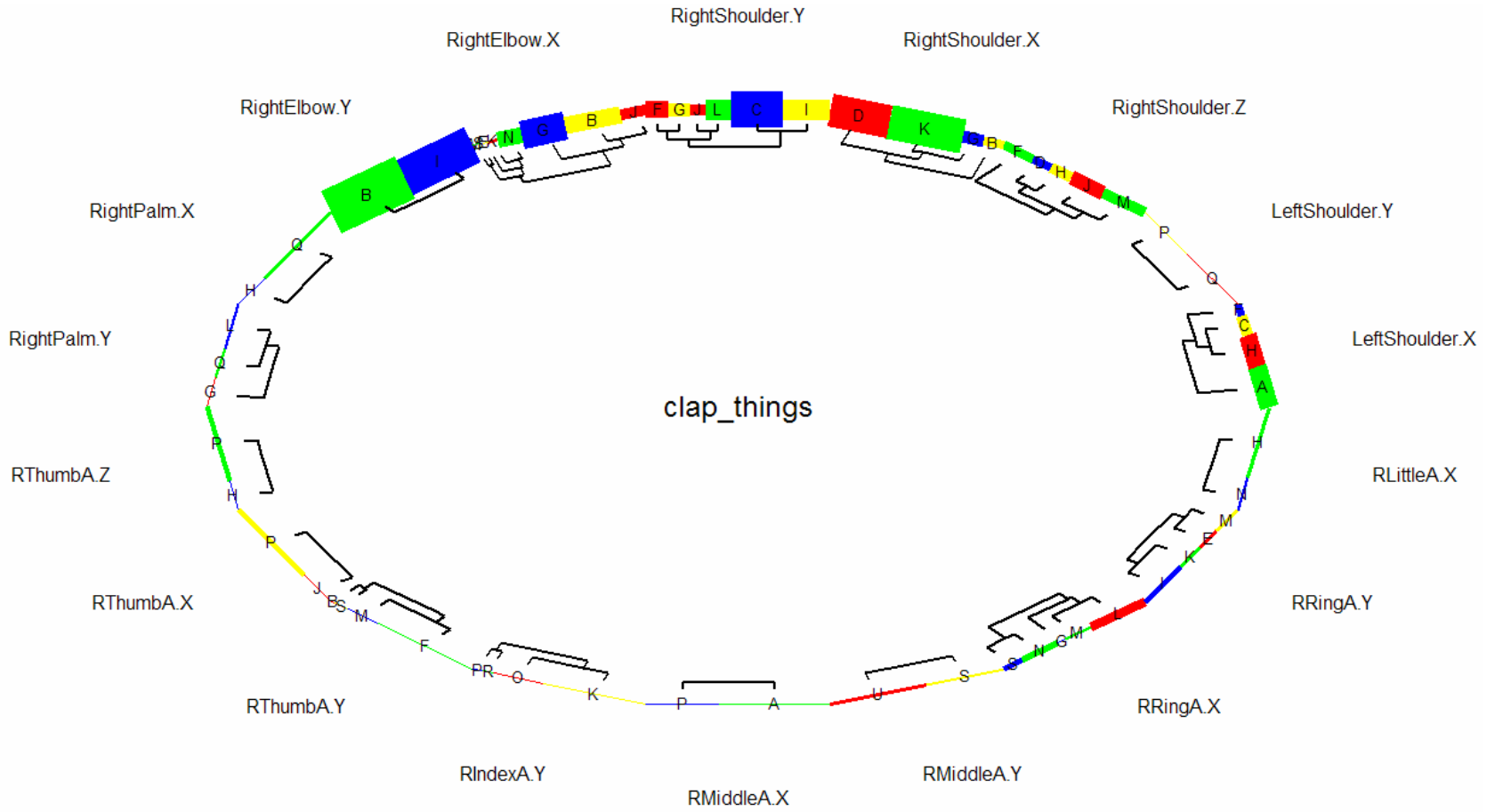
Appendix B: Words in HAL

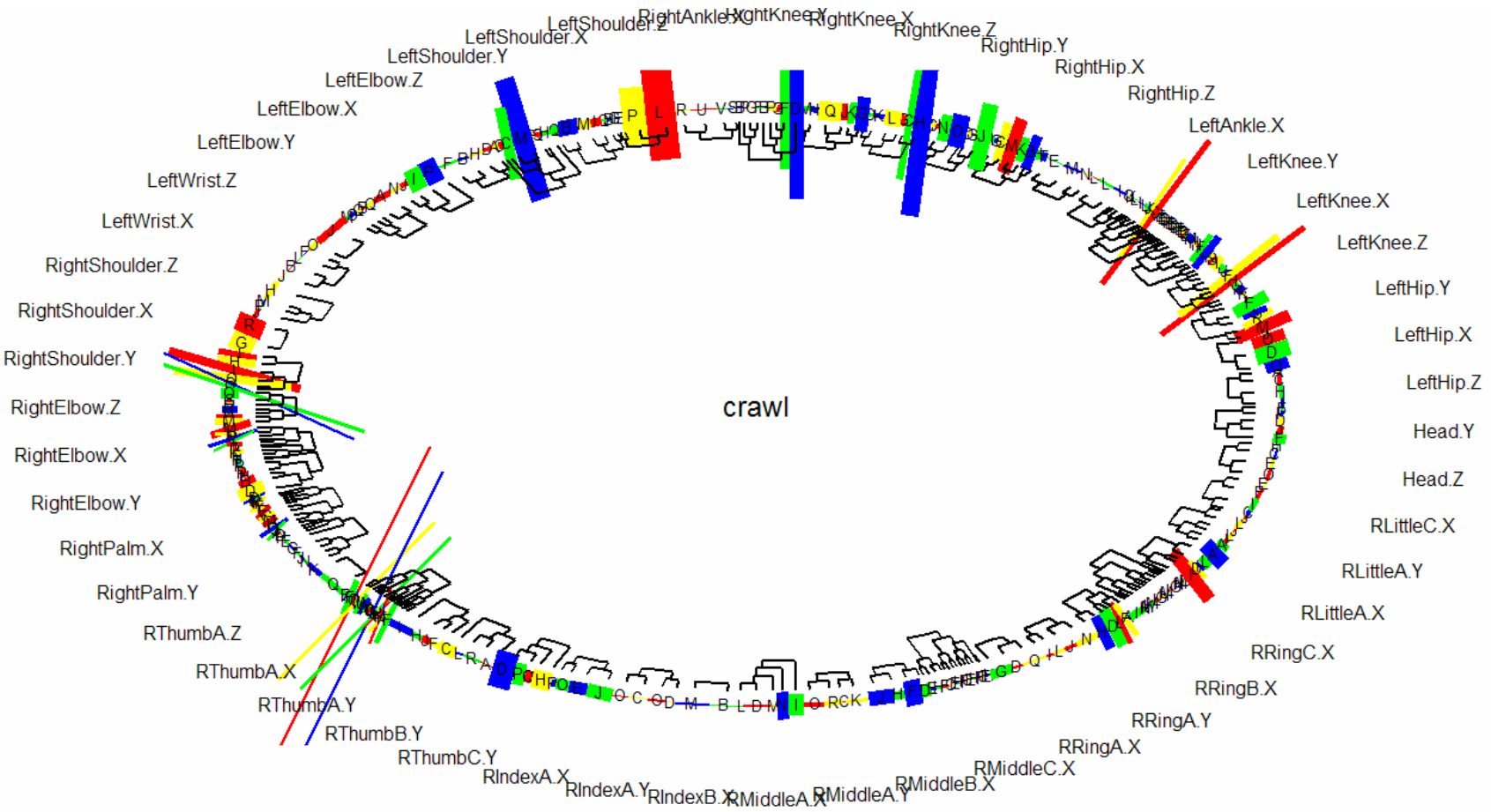


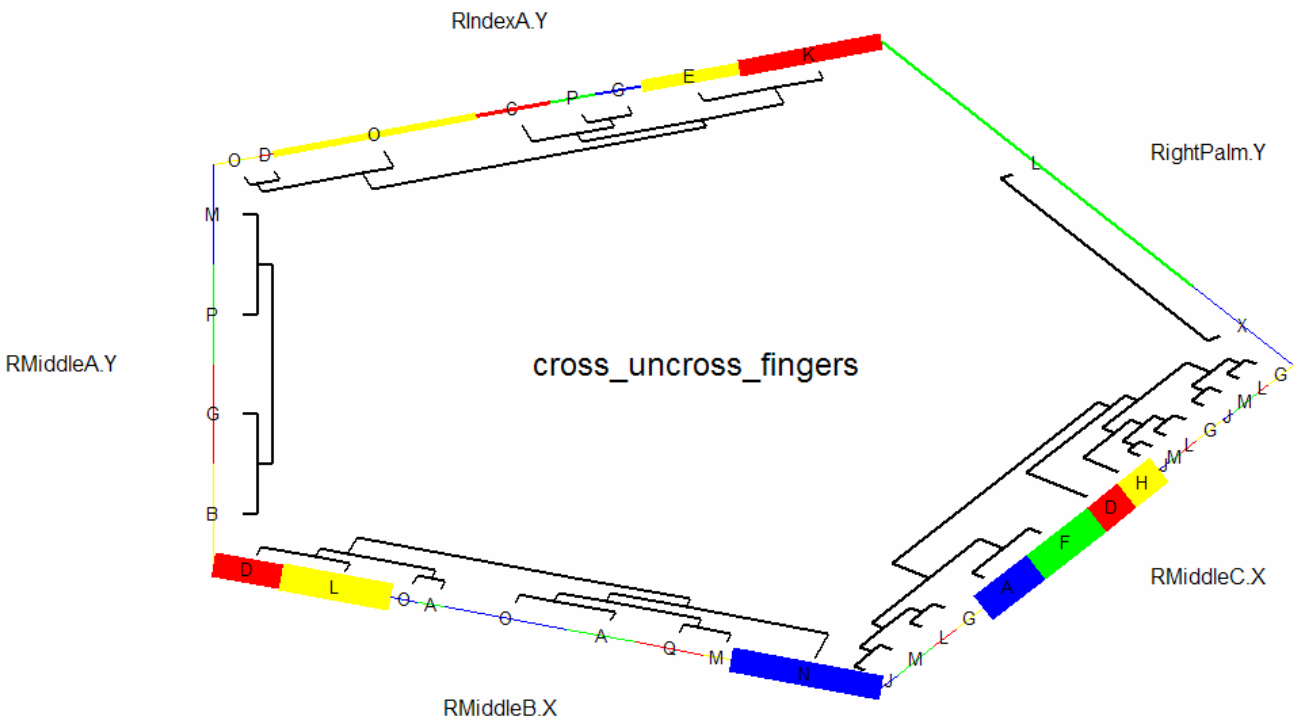

RightElbow.X
RightElbow.Z
RightElbow.Y
RightShoulder.Y
RightPalm.X
RightShoulder.X
RThumbA.Z
RightShoulder.Z
brush_hor
RThumbA.X
RMiddleB.X
RThumbA.Y
RMiddleA.X
RThumbC.Y
RIndexB.X
RIndexA.X
RIndexA.Y

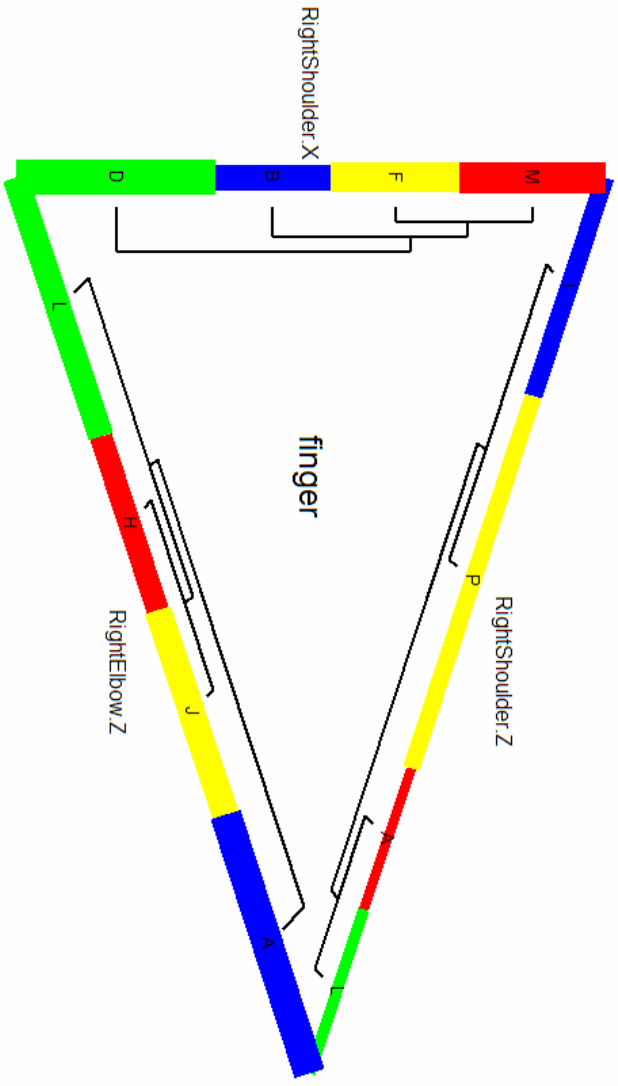

RightElbow.Z
RightElbow.X
RightShoulder.Y
RightElbow.Y
RightShoulder.X
RightPalm.Z
RightShoulder.Z
brush_ver
RightPalm.X
RIndexB.X
RThumbA.Z
RIndexA.Y
RThumbA.X
RThumbC.Y
RThumbA.Y

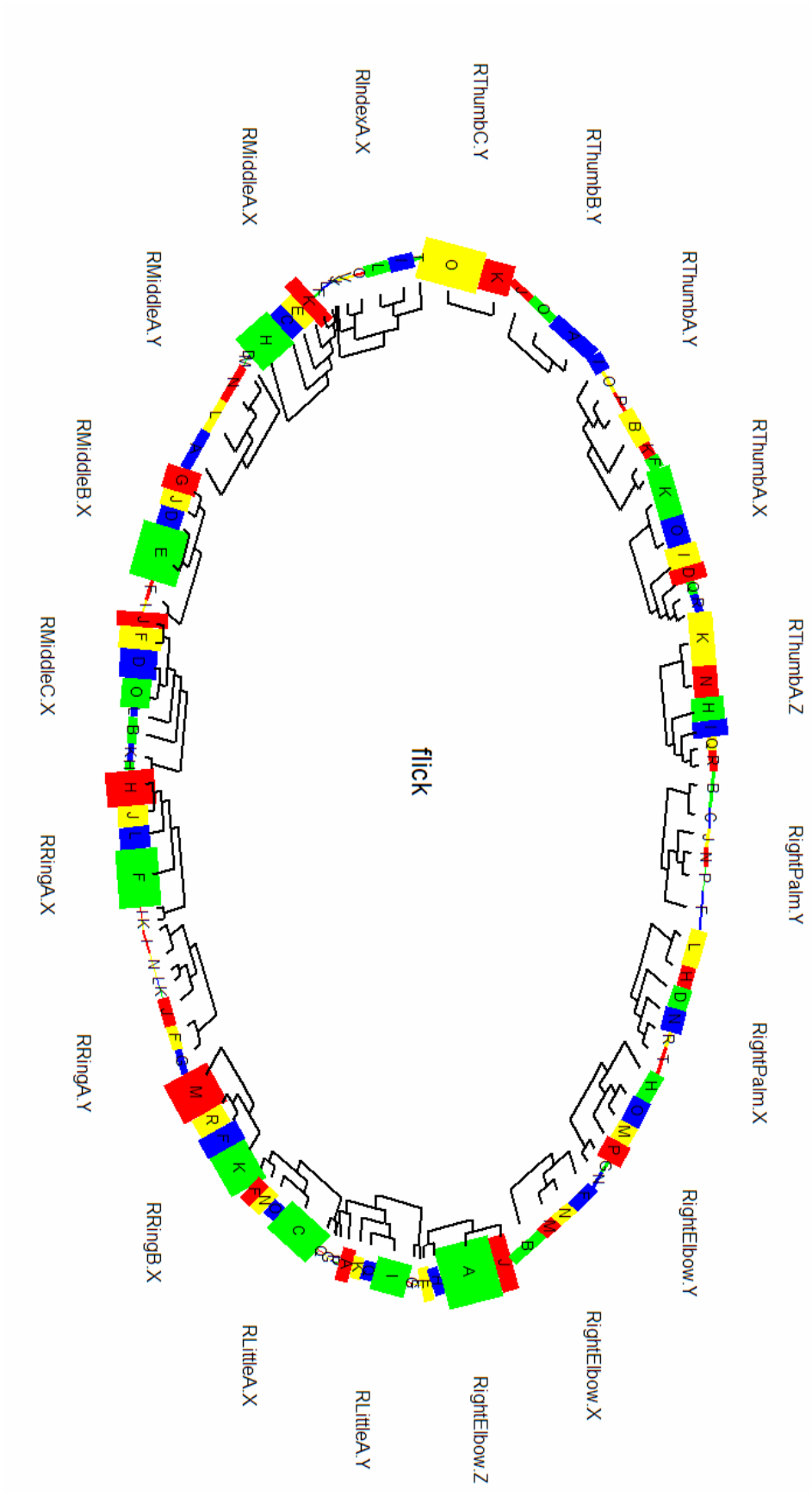

RightShoulder.X
RightShoulder.Y
RightShoulder.Z
RightElbow.Z
chop
RightPalm.Y
RightElbow.X
RightElbow.Y

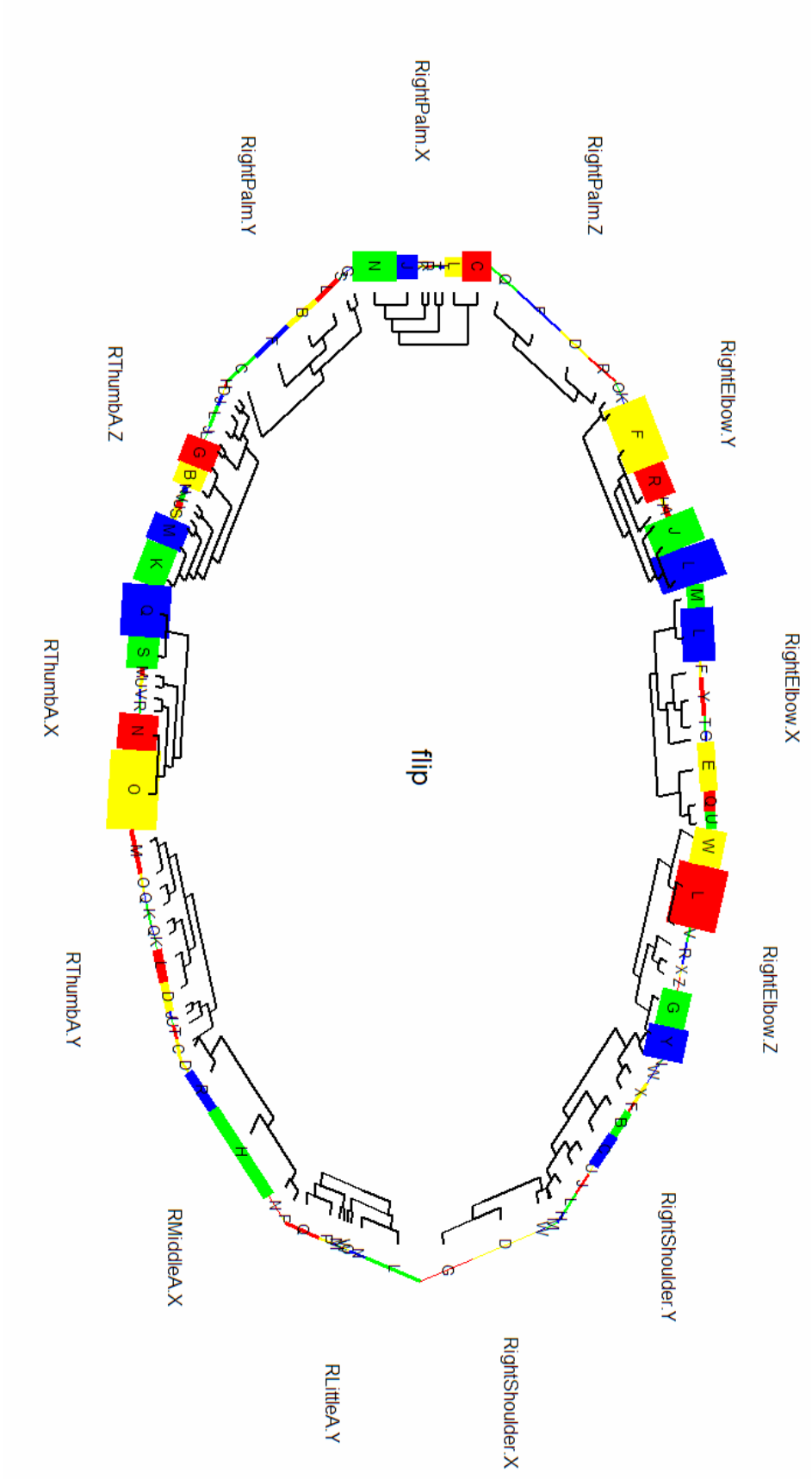

RightShoulder.Y
RightElbow.X
RightShoulder.X
RightElbow.Y
RightShoulder.Z
RightPalm.X
LeftShoulder.Y
RightPalm.Y
LeftShoulder.X
clap_things
RThumbA.Z
RLittleA.X
RThumbA.X
RRingA.Y
RThumbA.Y
RRingA.X
RIndexA.Y
RMiddleA.Y
RMiddleA.X

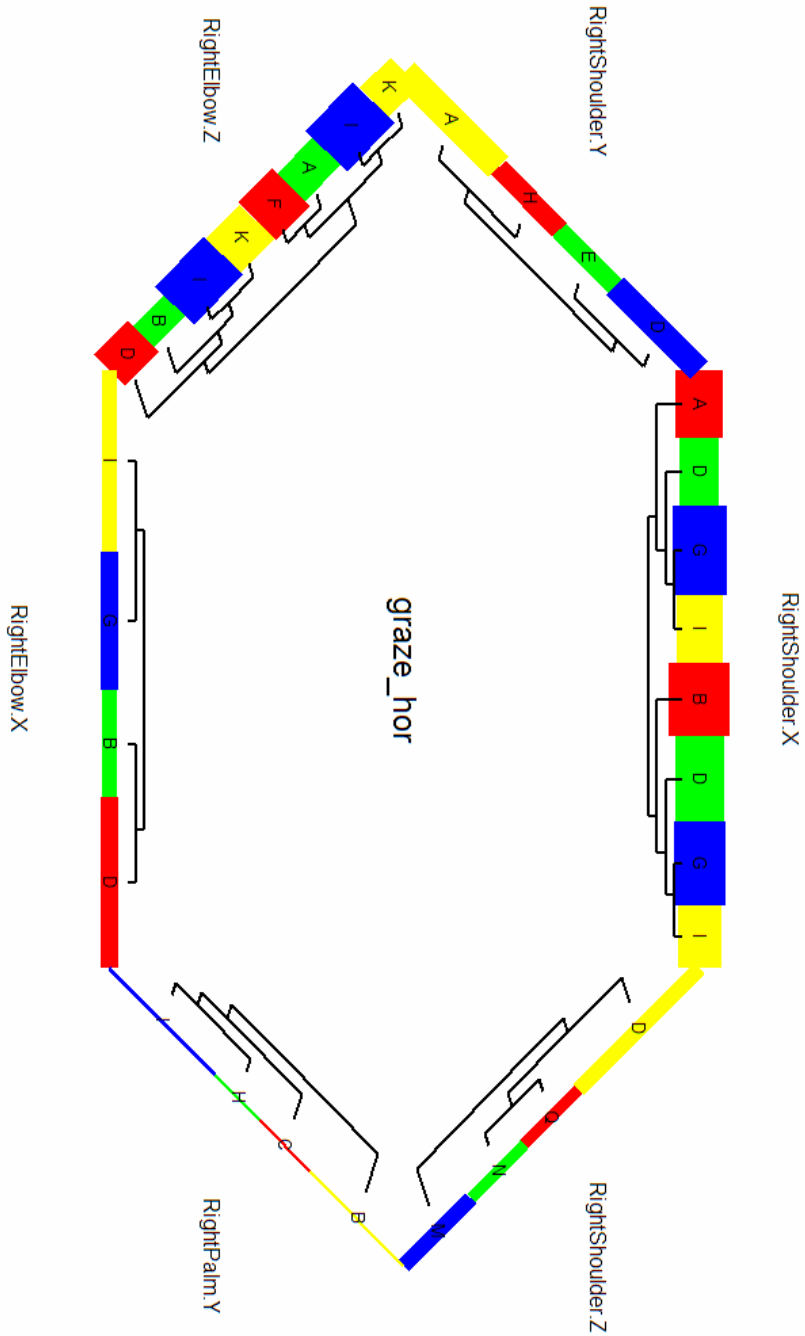

crawl
LeftShoulder.Z
LeftShoulder.X
LeftShoulder.Y
LeftElbow.Z
LeftElbow.X
LeftElbow.Y
LeftWrist.Z
LeftWrist.X
RightShoulder.Z
RightShoulder.X
RightShoulder.Y
RightElbow.Z
RightElbow.X
RightElbow.Y
RightPalm.X
RightPalm.Y
RThumbA.Z
RThumbA.X
RThumbA.Y
RThumbB.Y
RThumbC.Y
RIndexA.X
RIndexA.Y
RIndexB.X
RMiddleA.X
RMiddleA.Y
RMiddleB.X
RMiddleC.X
RRingA.X
RRingA.Y
RRingB.X
RRingC.X
RLittleA.X
RLittleA.Y
RLittleC.X
Head.Z
Head.Y
LeftHip.Z
LeftHip.X
LeftHip.Y
LeftKnee.Z
LeftKnee.X
LeftKnee.Y
LeftAnkle.X
RightHip.Z
RightHip.X
RightHip.Y
RightKnee.Z
RightKnee.X
RightAnkle.X

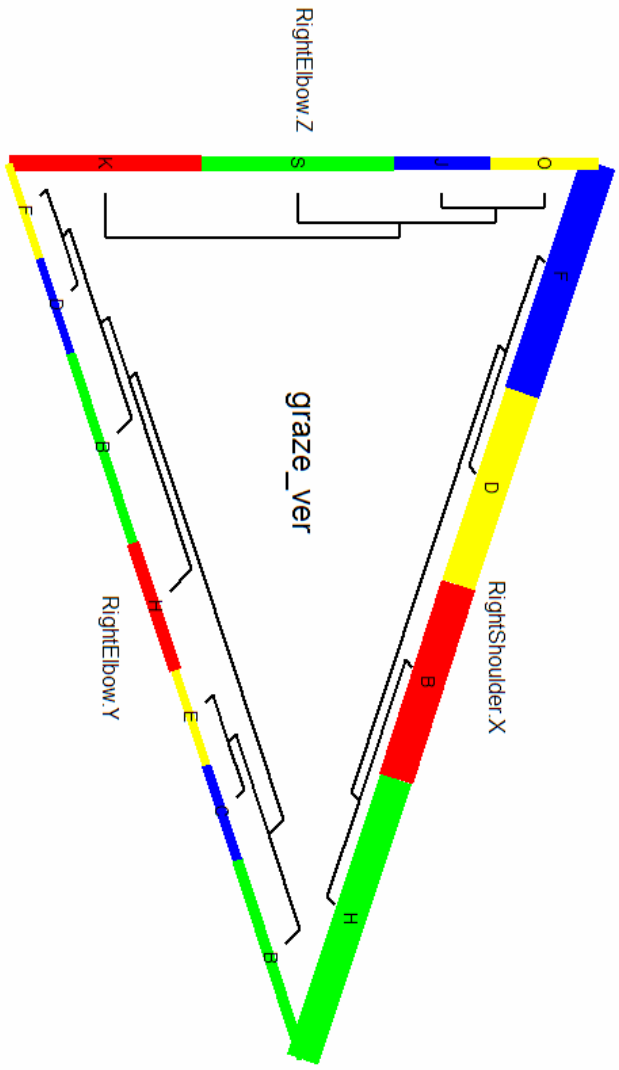

RIndexA.Y
RightPalm.Y
RMiddleA.Y
cross_uncross_fingers
RMiddleC.X
RMiddleB.X

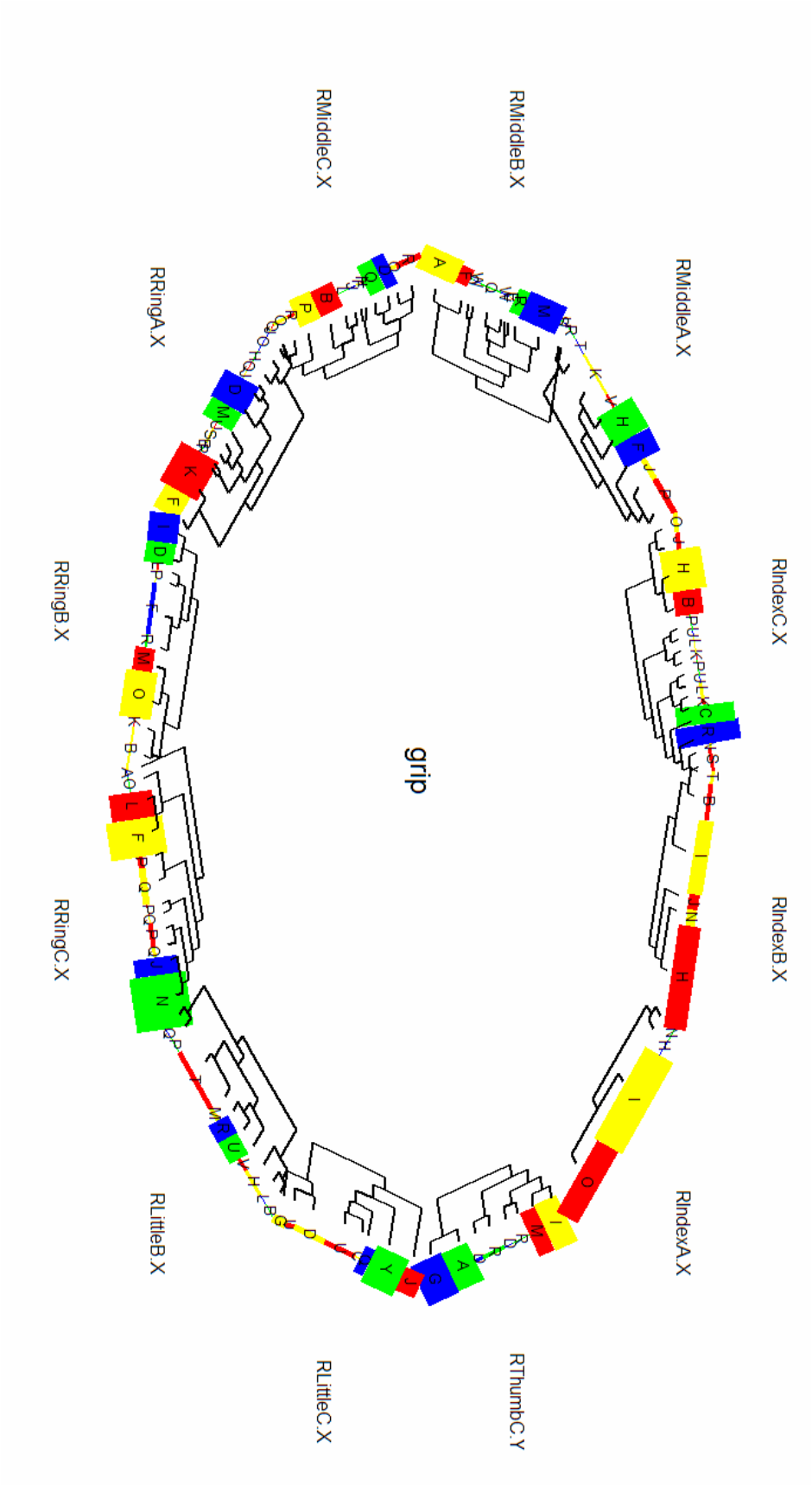

RightShoulder.X
D
B
F
M
L
finger
H
RightElbow.Z
J
P
RightShoulder.Z
A
A

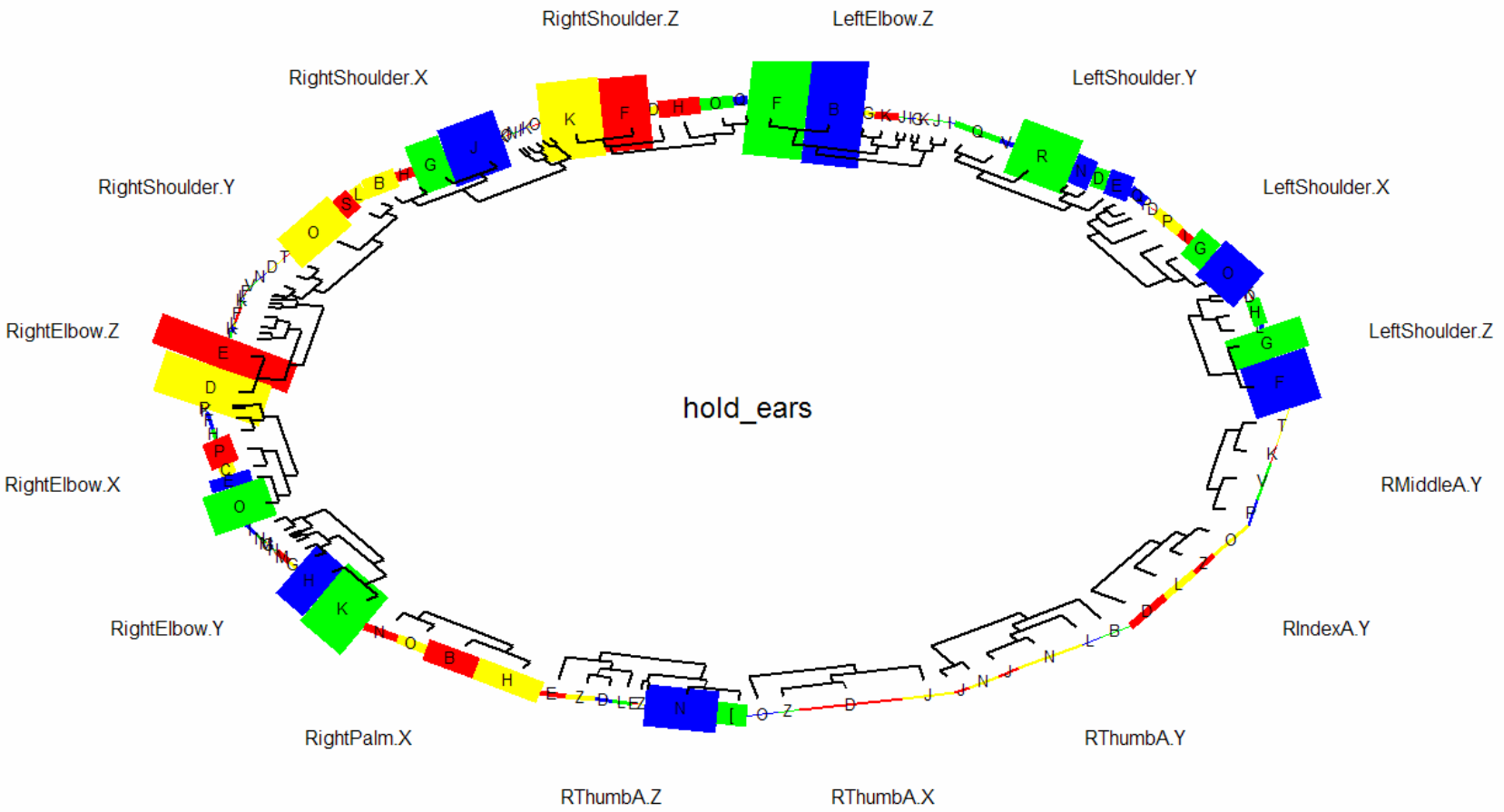

flick
RThumbA.Z
RightPalm.Y
RightPalm.X
RightElbow.Y
RightElbow.X
RightElbow.Z
RLittleA.Y
RLittleA.X
RRingB.X
RRingA.Y
RRingA.X
RMiddleC.X
RMiddleB.X
RMiddleA.Y
RMiddleA.X
RIndexA.X
RThumbC.Y
RThumbB.Y
RThumbA.Y
RThumbA.X

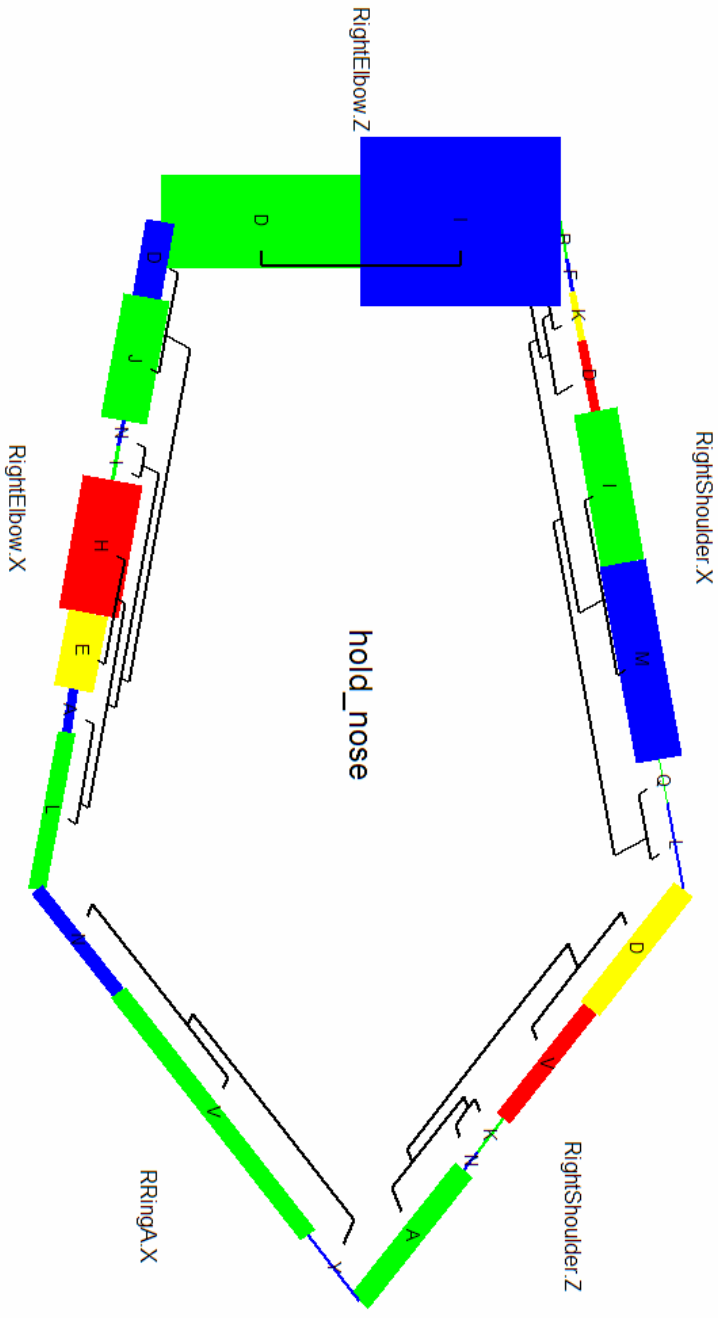

RightPalm.X
RightPalm.Z
RightElbow.Y
RightElbow.X
RightElbow.Z
RightShoulder.Y
RightShoulder.X
RLittleA.Y
RMiddleA.X
RThumbA.Y
RThumbA.X
RThumbA.Z
RightPalm.Y
flip

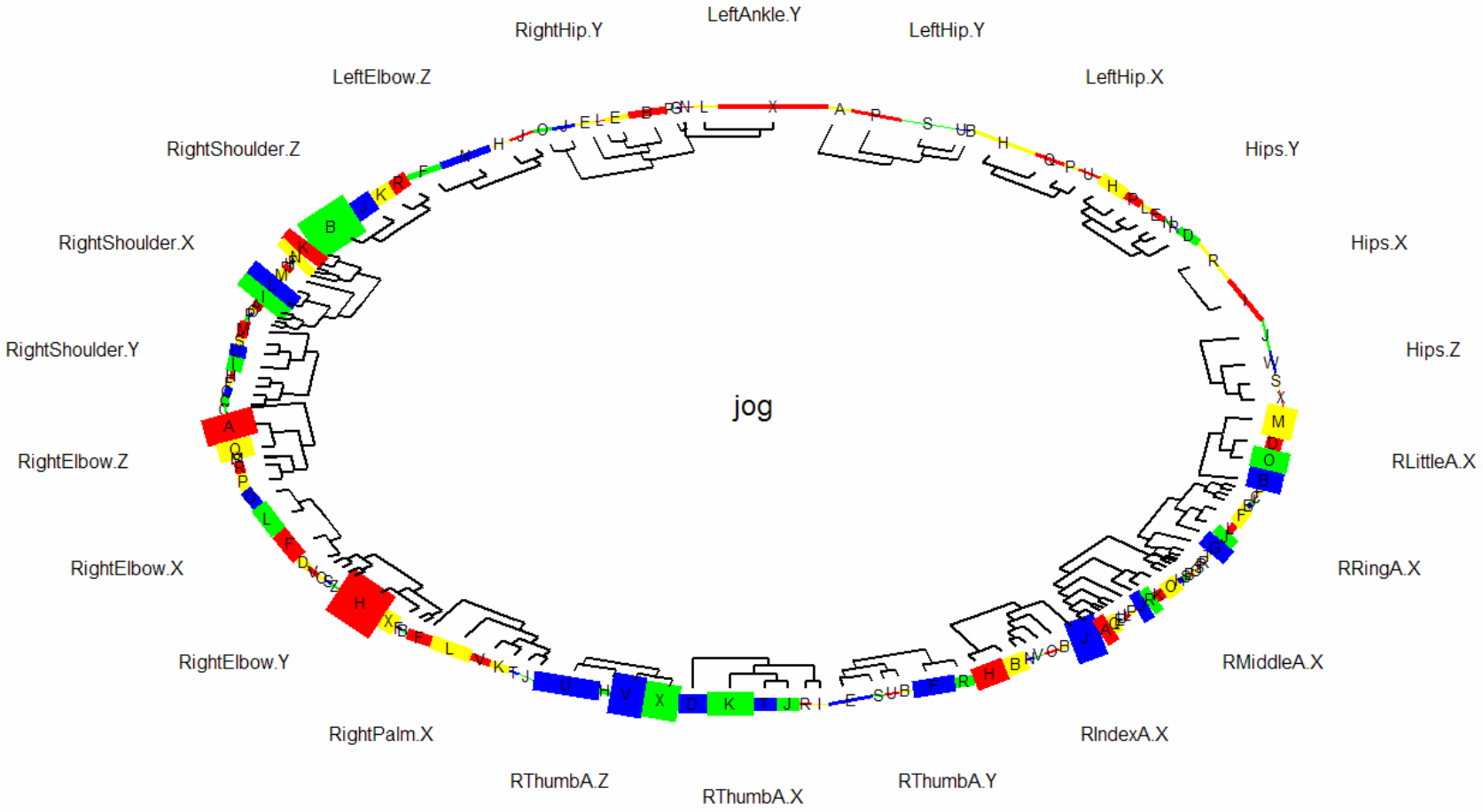

graze_hor
RightShoulder.Y
RightShoulder.X
RightShoulder.Z
RightPalm.Y
RightElbow.X
RightElbow.Z

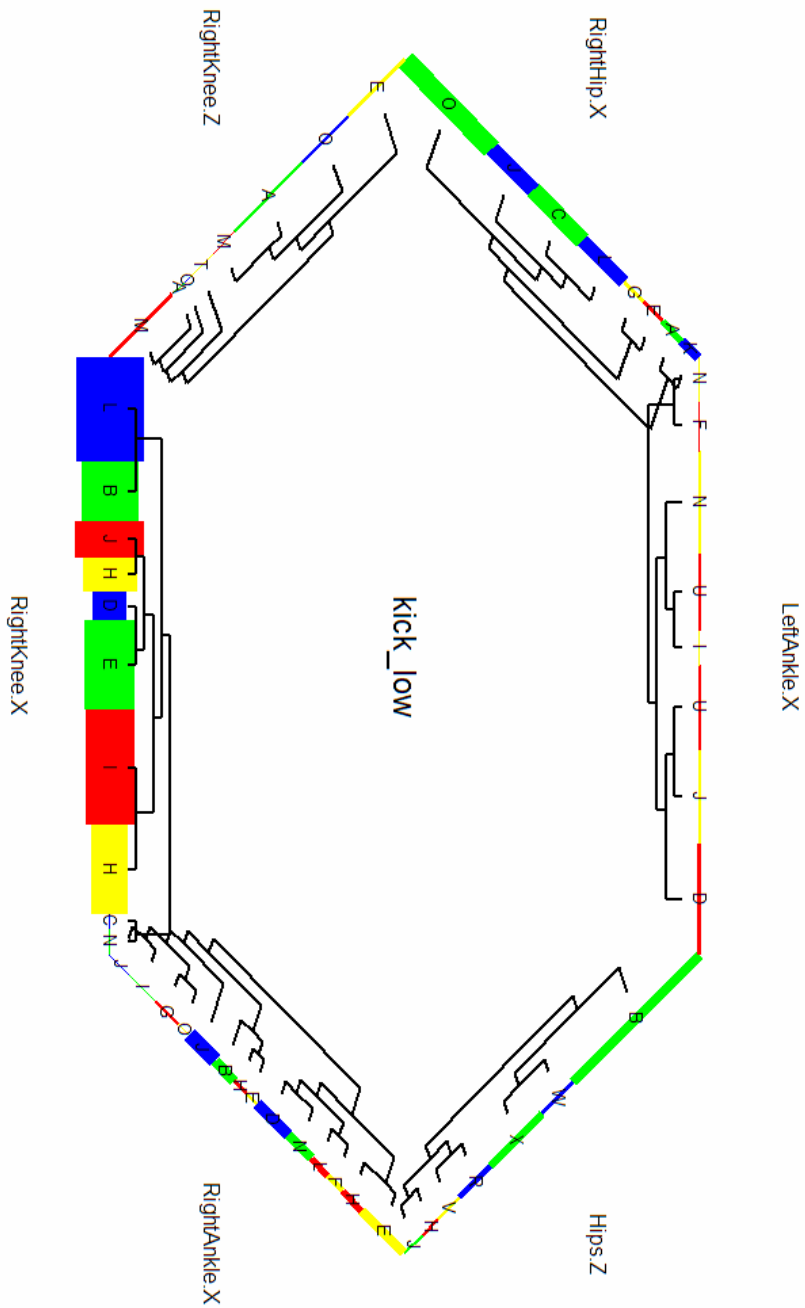

RightElbow.Z
graze_ver
RightShoulder.X
RightElbow.Y

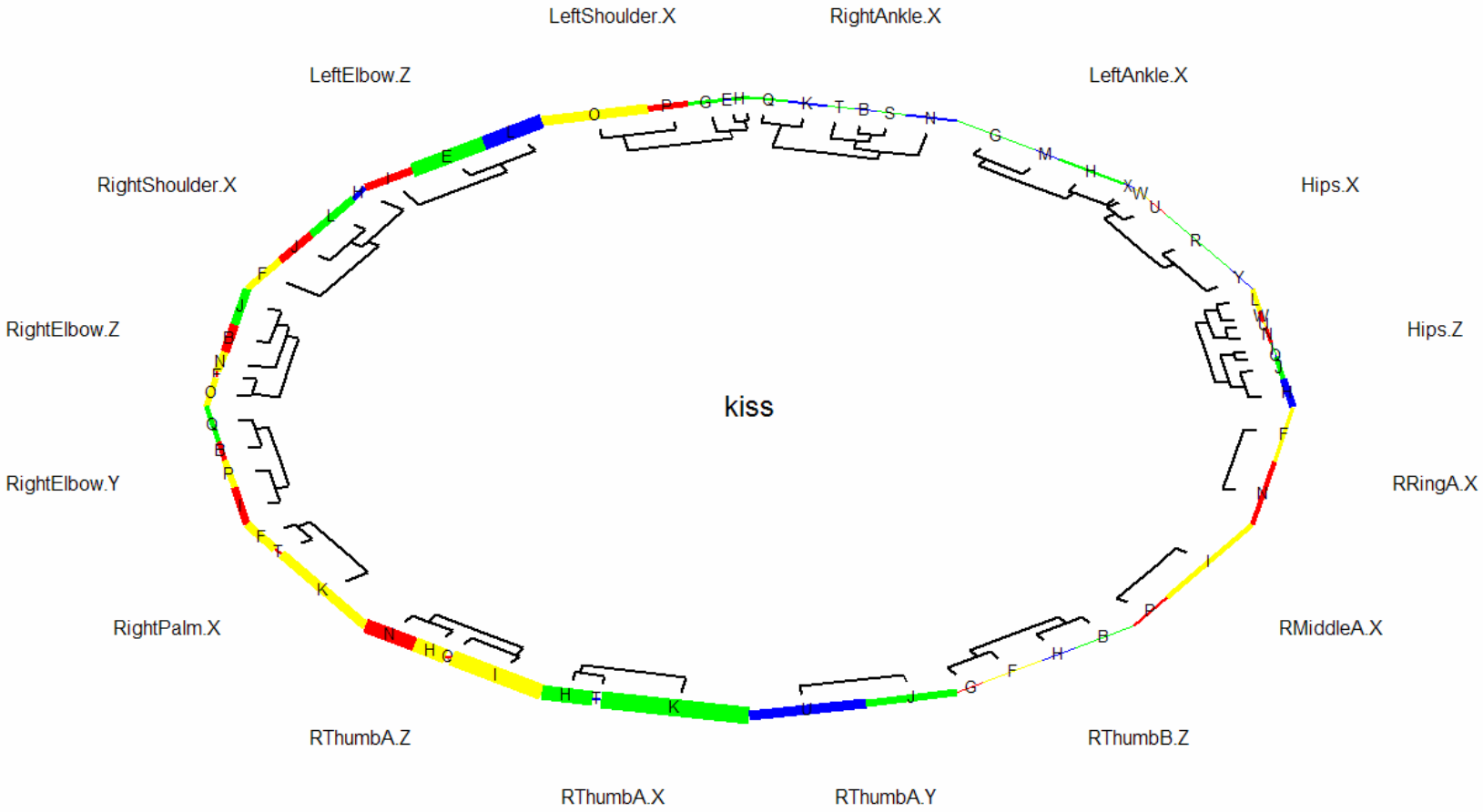

RMiddleC.X
RMiddleB.X
RMiddleA.X
RIndexC.X
RIndexB.X
RIndexA.X
RThumbC.Y
RLittleC.X
RLittleB.X
RRingC.X
RRingB.X
RRingA.X
grip

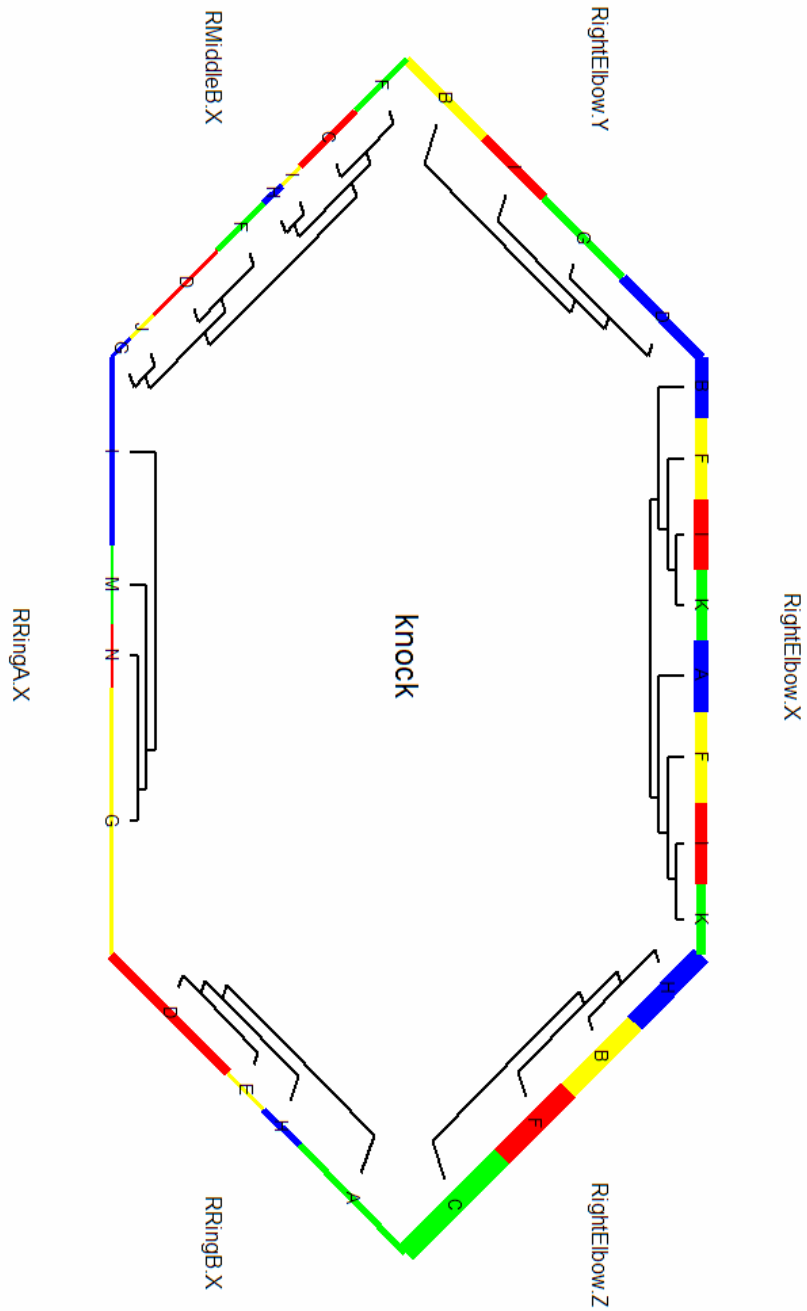

RightShoulder.Z
LeftElbow.Z
RightShoulder.X
LeftShoulder.Y
RightShoulder.Y
LeftShoulder.X
RightElbow.Z
LeftShoulder.Z
hold_ears
RightElbow.X
RMiddleA.Y
RightElbow.Y
RIndexA.Y
RightPalm.X
RThumbA.Y
RThumbA.Z
RThumbA.X

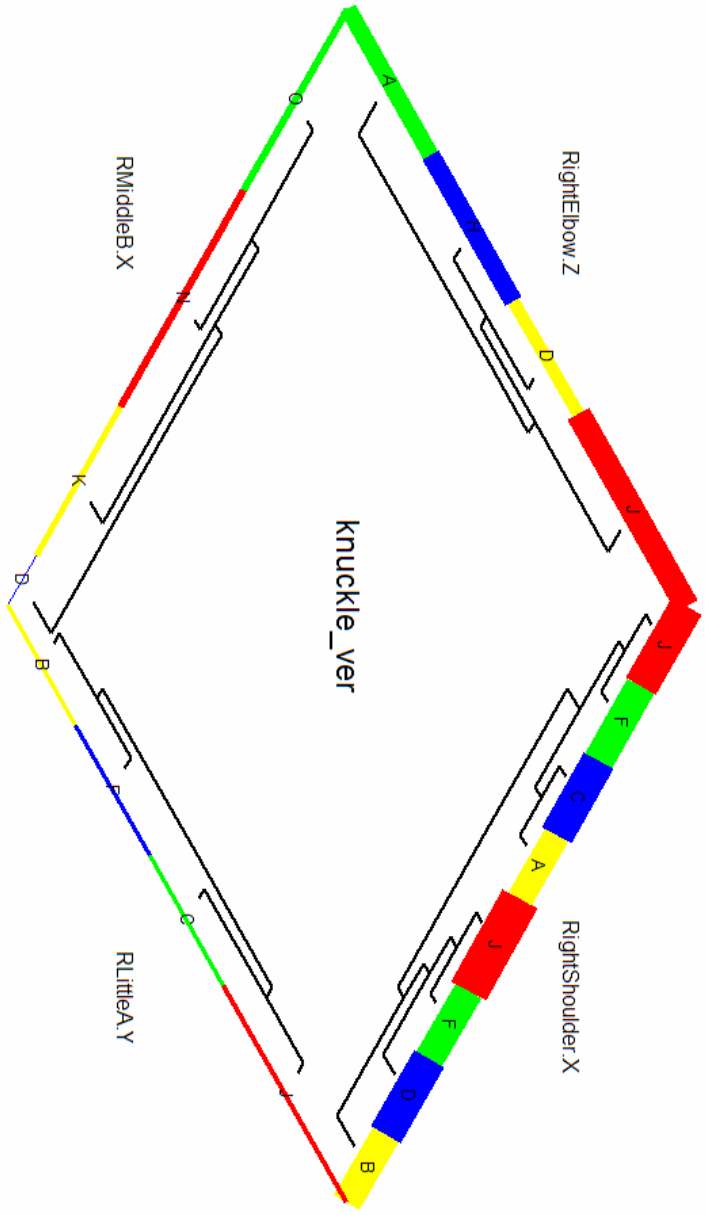

RightElbow.Z
RightShoulder.X
RightElbow.X
hold_nose
RRingA.X
RightShoulder.Z


LeftAnkle.Y
RightHip.Y
LeftHip.Y
LeftElbow.Z
LeftHip.X
RightShoulder.Z
Hips.Y
RightShoulder.X
Hips.X
RightShoulder.Y
Hips.Z
jog
RightElbow.Z
RLittleA.X
RightElbow.X
RRingA.X
RightElbow.Y
RMiddleA.X
RightPalm.X
RIndexA.X
RThumbA.Z
RThumbA.X
RThumbA.Y

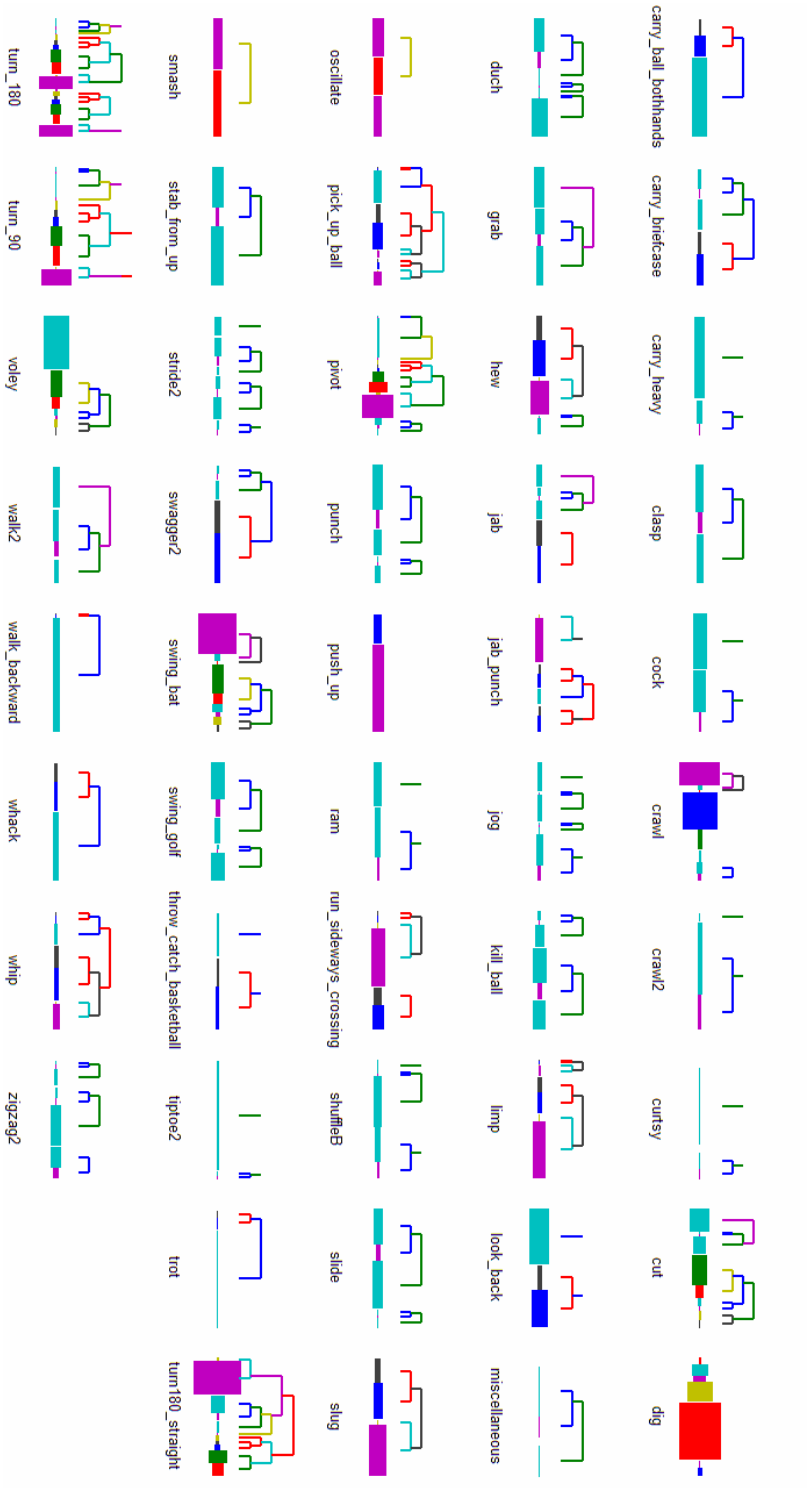

RightKnee.Z
RightHip.X
RightKnee.X
kick_low
LeftAnkle.X
RightAnkle.X
Hips.Z


LeftShoulder.X
RightAnkle.X
LeftElbow.Z
LeftAnkle.X
RightShoulder.X
Hips.X
RightElbow.Z
Hips.Z
kiss
RightElbow.Y
RRingA.X
RightPalm.X
RMiddleA.X
RThumbA.Z
RThumbB.Z
RThumbA.X
RThumbA.Y

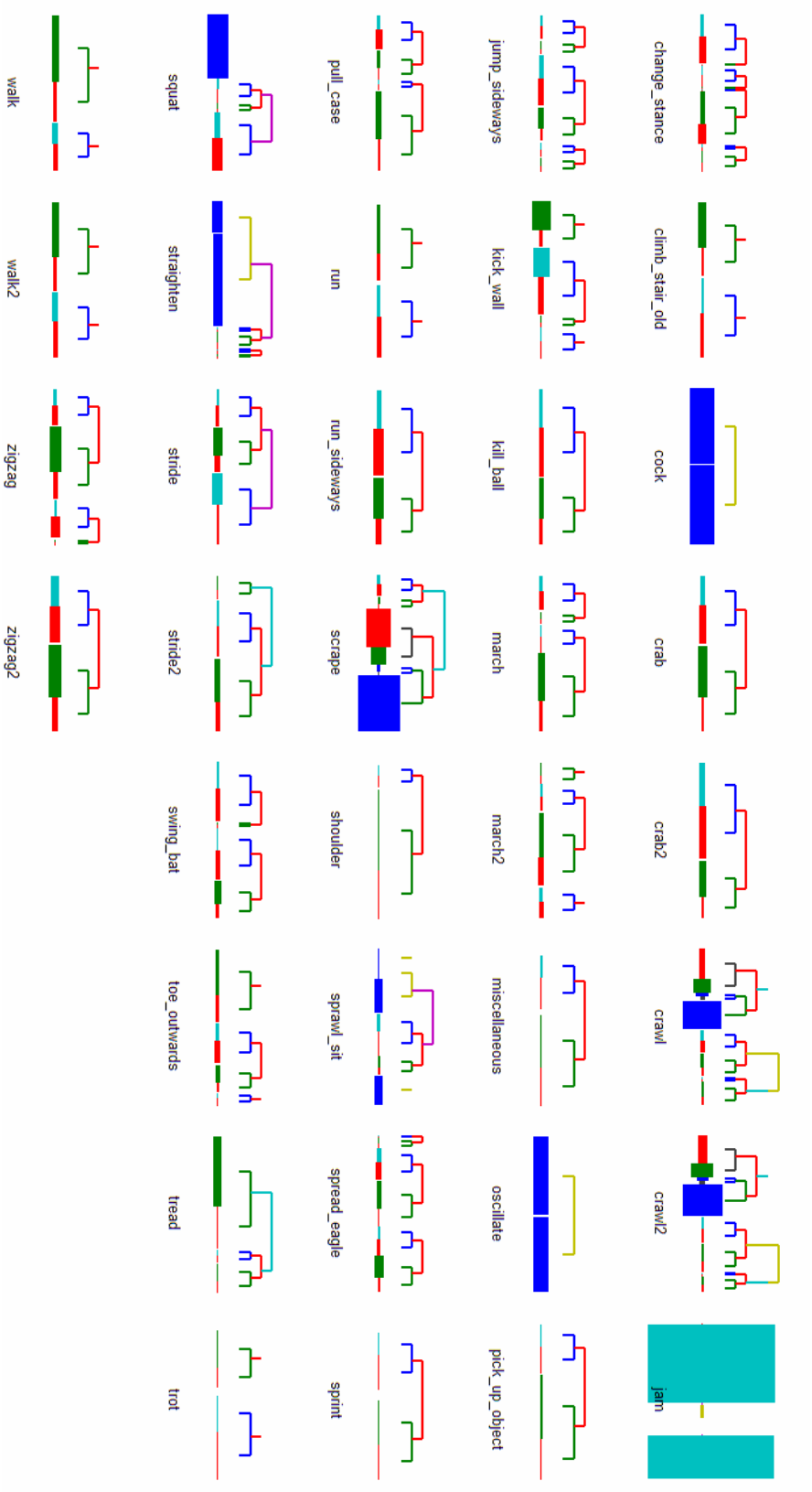

RightElbow.Y
RightElbow.X
RightElbow.Z
RRingB.X
RRingA.X
RMiddleB.X
knock


knuckle_ver
RightElbow.Z
RightShoulder.X
RLittleA.Y
RMiddleB.X

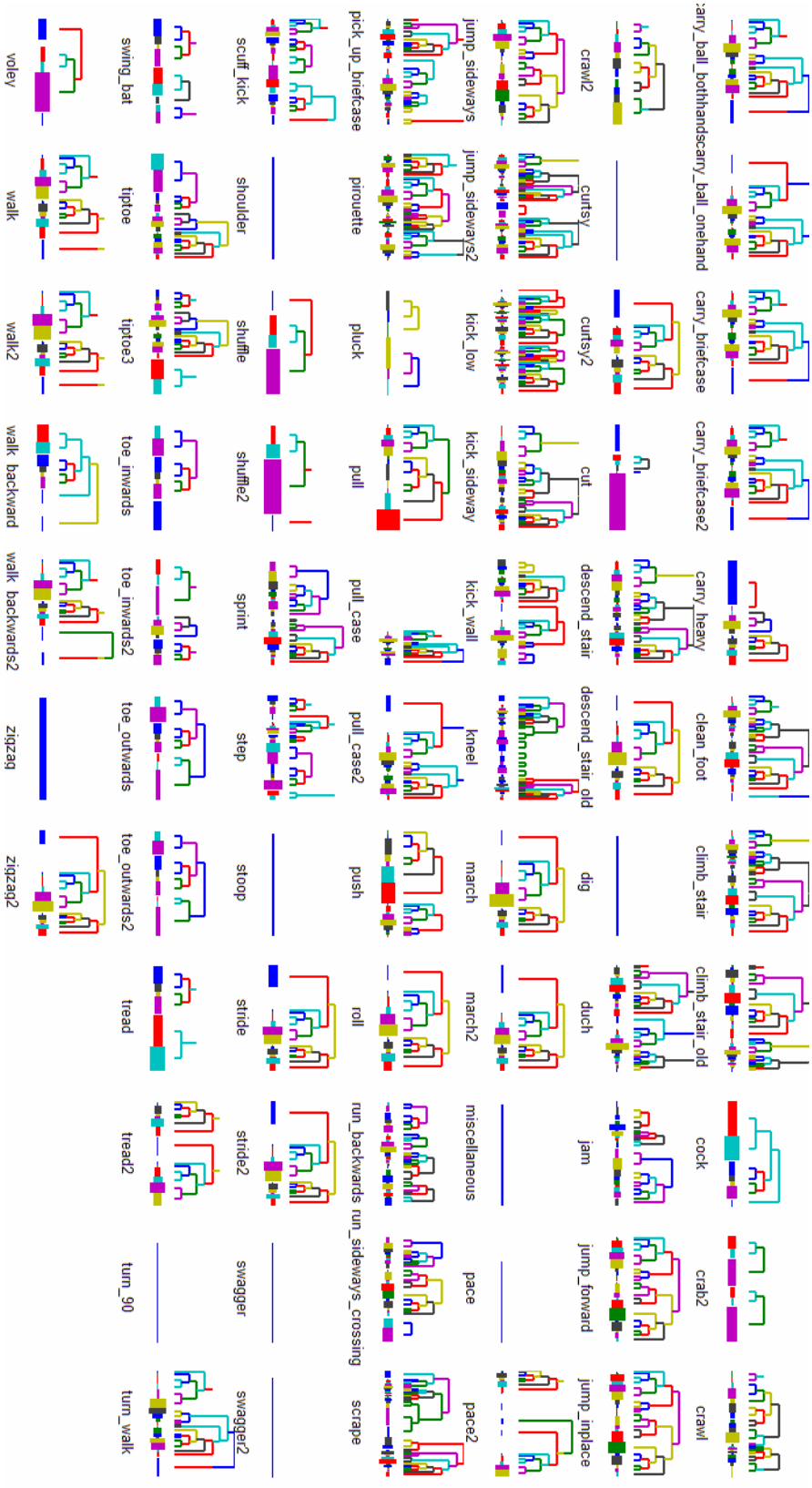
Appendix C: Morphological Grammars



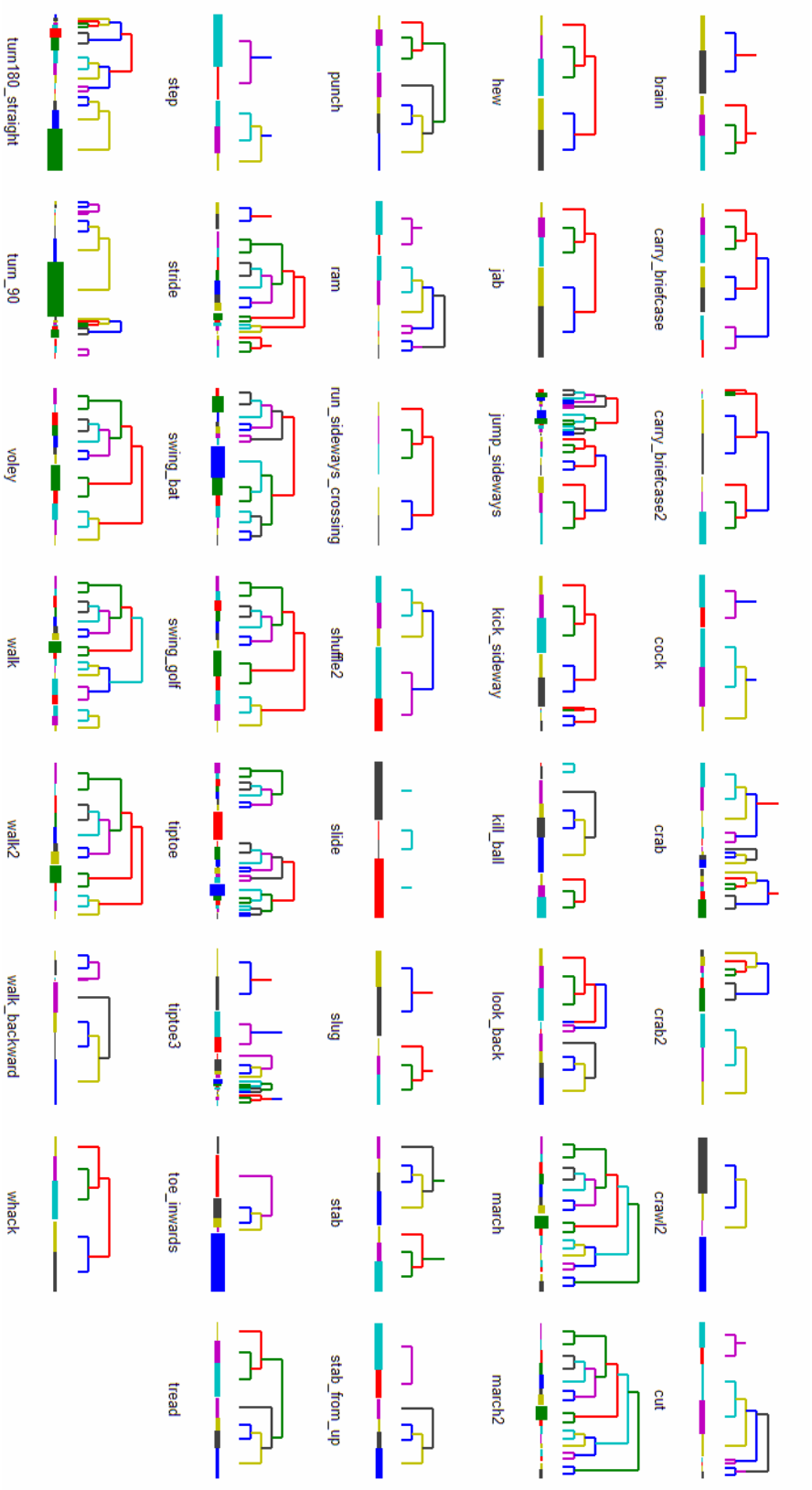
Left Hip Y


change_stance
climb_stair_old
cock
crab
crab2
crawl
crawl2
jam
jump_sideways
kick_wall
kill_ball
march
march2
miscellaneous
oscillate
pick_up_object
pull_case
run
run_sideways
scrape
shoulder
sprawl_sit
spread_eagle
sprint
squat
straighten
stride
stride2
swing_bat
toe_outwards
tread
trot
walk
walk2
zigzag
zigzag2

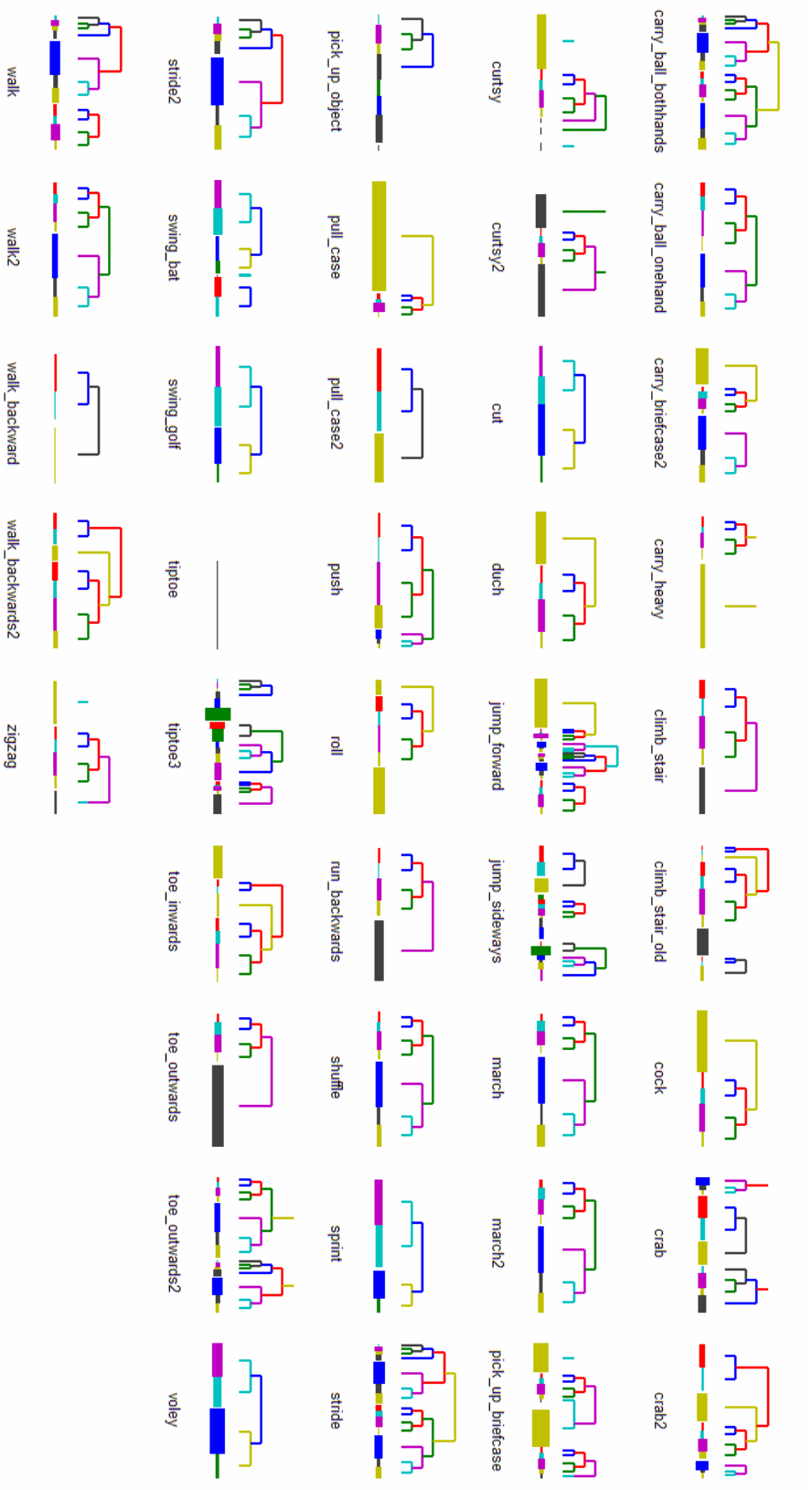
Right Hip Z


carry_ball_bothhandscarry_ball_onehand
crawl2
curtsy
jump_sideways
jump_sideways2
pick_up_briefcase
pirouette
scuff_kick
shoulder
swing_bat
tiptoe
voley
walk
carry_briefcase
curtsy2
kick_low
pluck
walk2
tiptoe3
carry_briefcase2
cut
kick_sideway
pull
shuffle
shuffle2
toe_inwards
walk_backward
carry_heavy
descend_stair
kick_wall
pull_case
sprint
toe_inwards2
walk_backwards2
clean_foot
descend_stair_old
kneel
pull_case2
step
toe_outwards
zigzag
climb_stair
dig
march
push
stoop
toe_outwards2
zigzag2
climb_stair_old
duch
march2
roll
stride
tread
cock
jam
miscellaneous
pace
run_backwards
stride2
tread2
crab2
jump_forward
pace2
run_sideways_crossing
swagger
turn_90
crawl
jump_inplace
scrape
swagger2
turn_walk

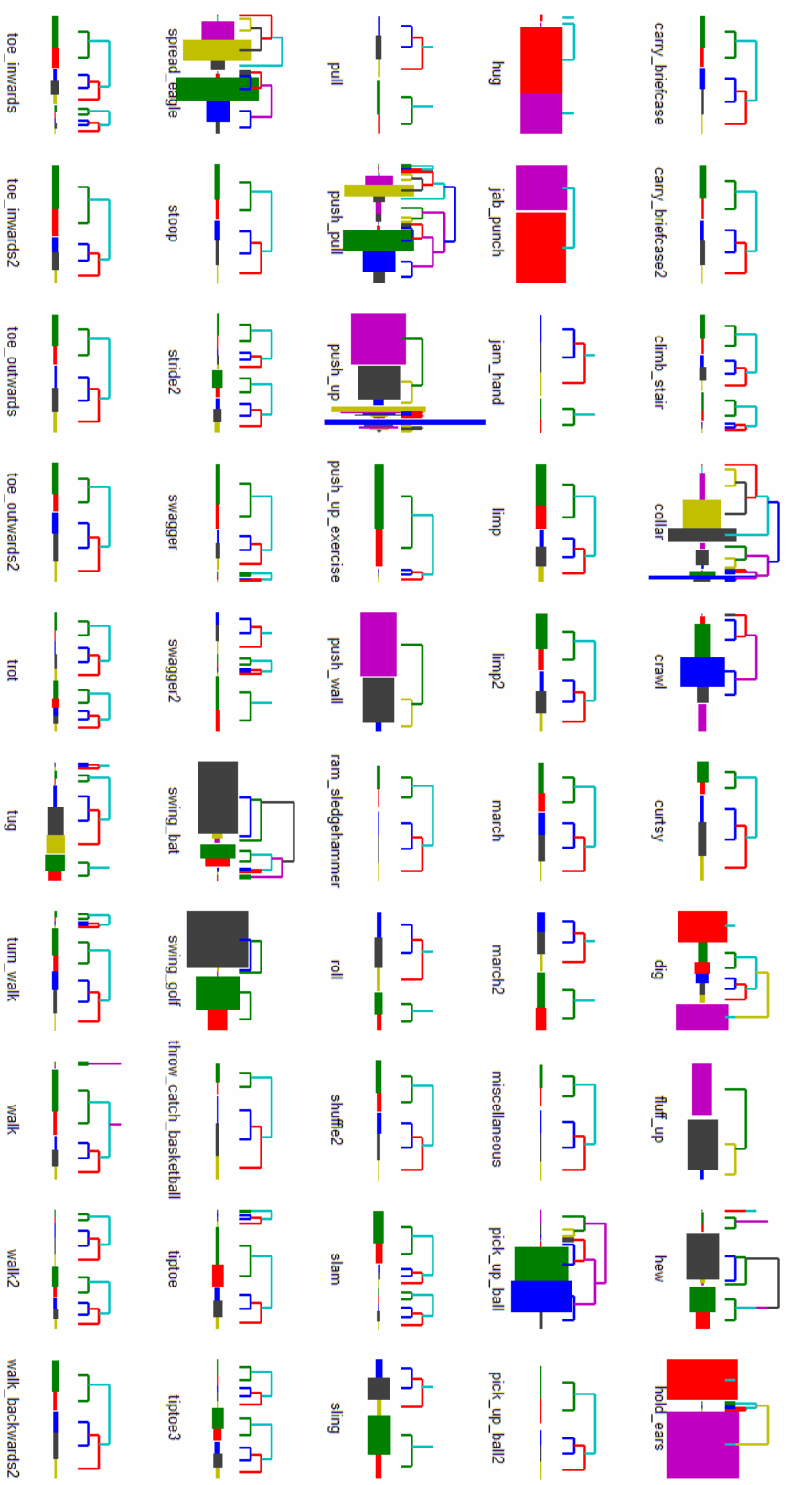
Right Knee X


brain
carry_briefcase
carry_briefcase2
cock
crab
crab2
crawl2
cut
hew
jab
jump_sideways
kick_sideway
kill_ball
look_back
march
march2
punch
ram
run_sideways_crossing
shuffle2
slide
slug
stab
stab_from_up
step
stride
swing_bat
swing_golf
tiptoe
tiptoe3
toe_inwards
tread
turn180_straight
turn_90
voley
walk
walk2
walk_backward
whack

Left Ankle Z

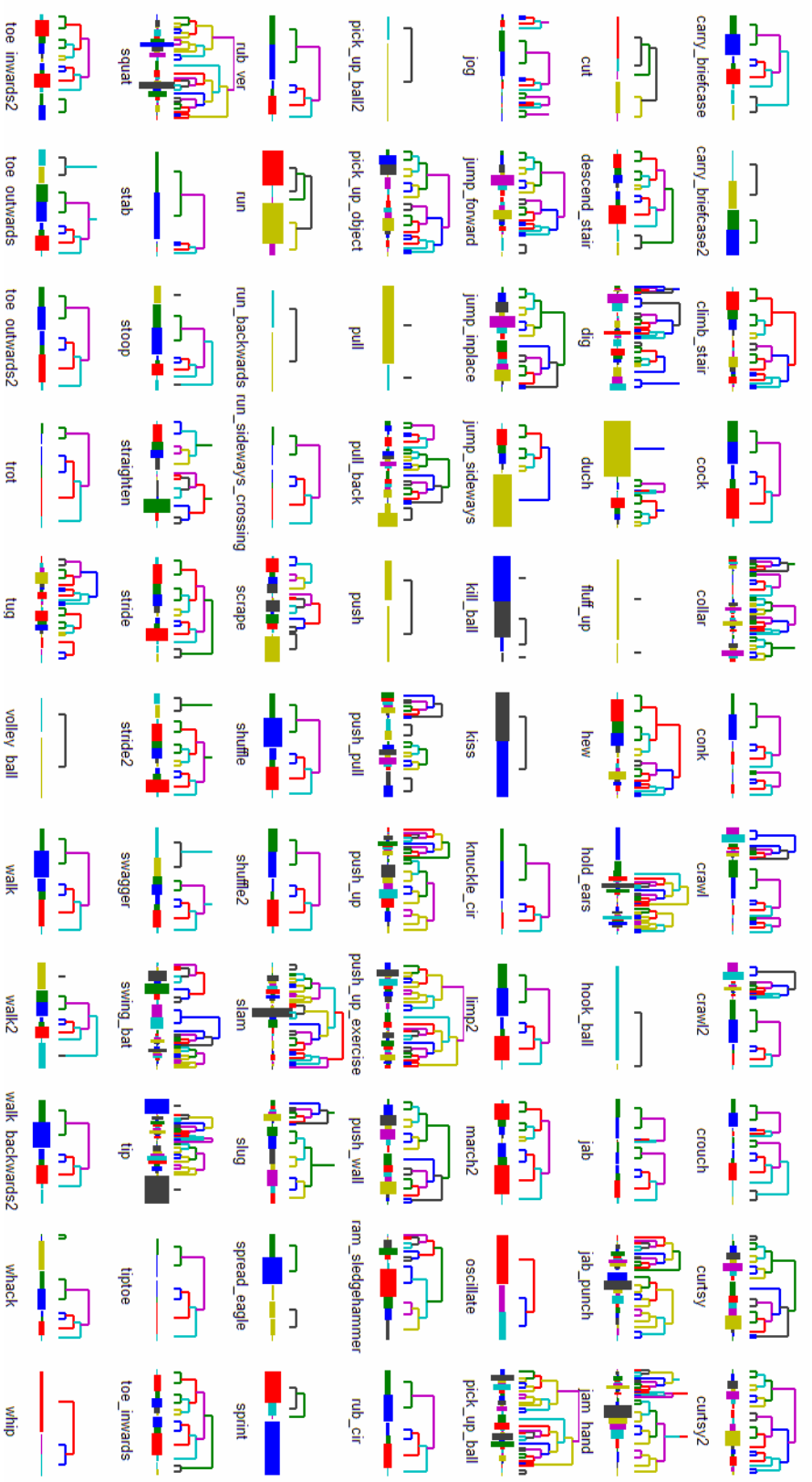


Right Ankle Y

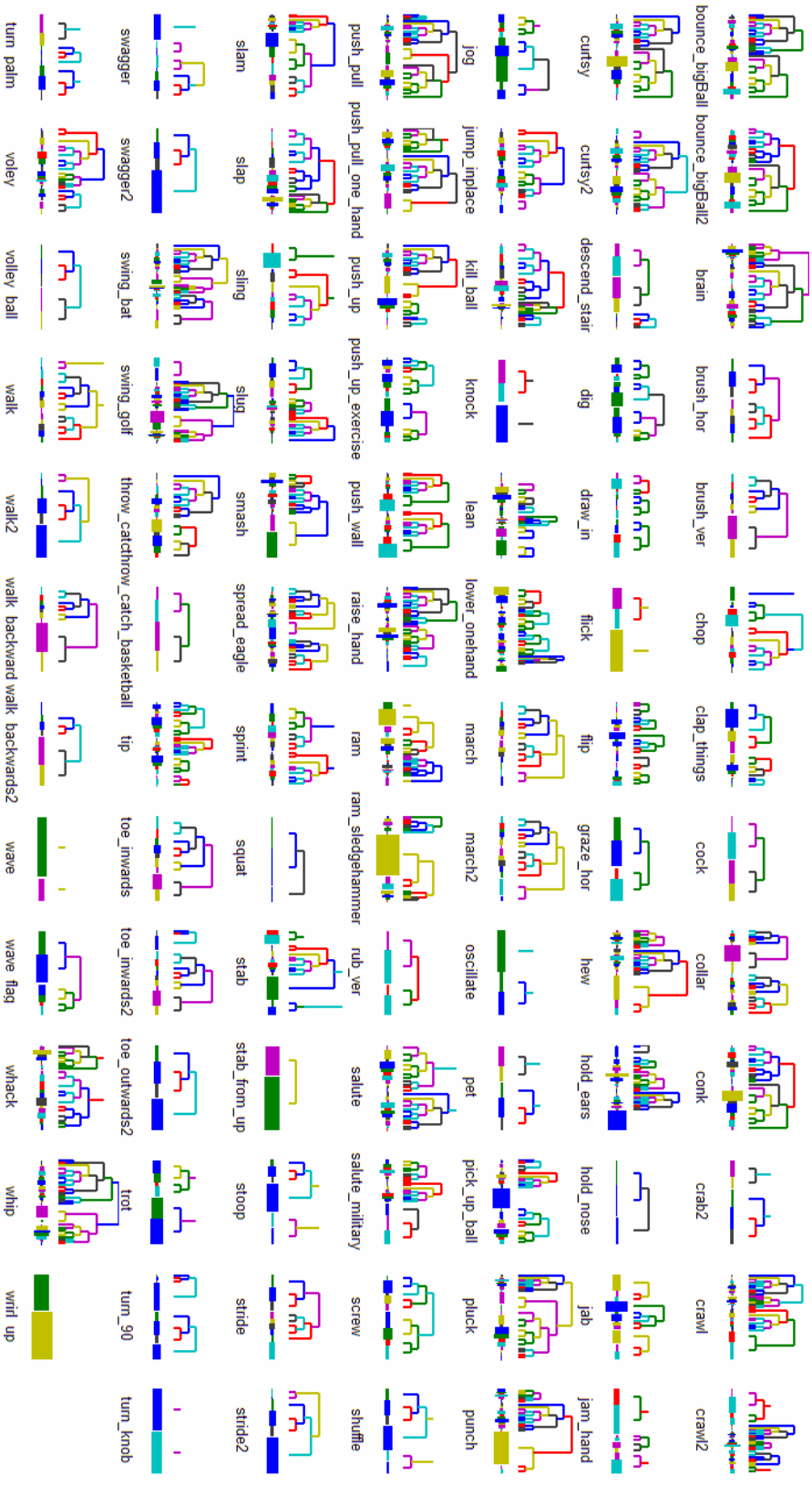


Left Shoulder Z

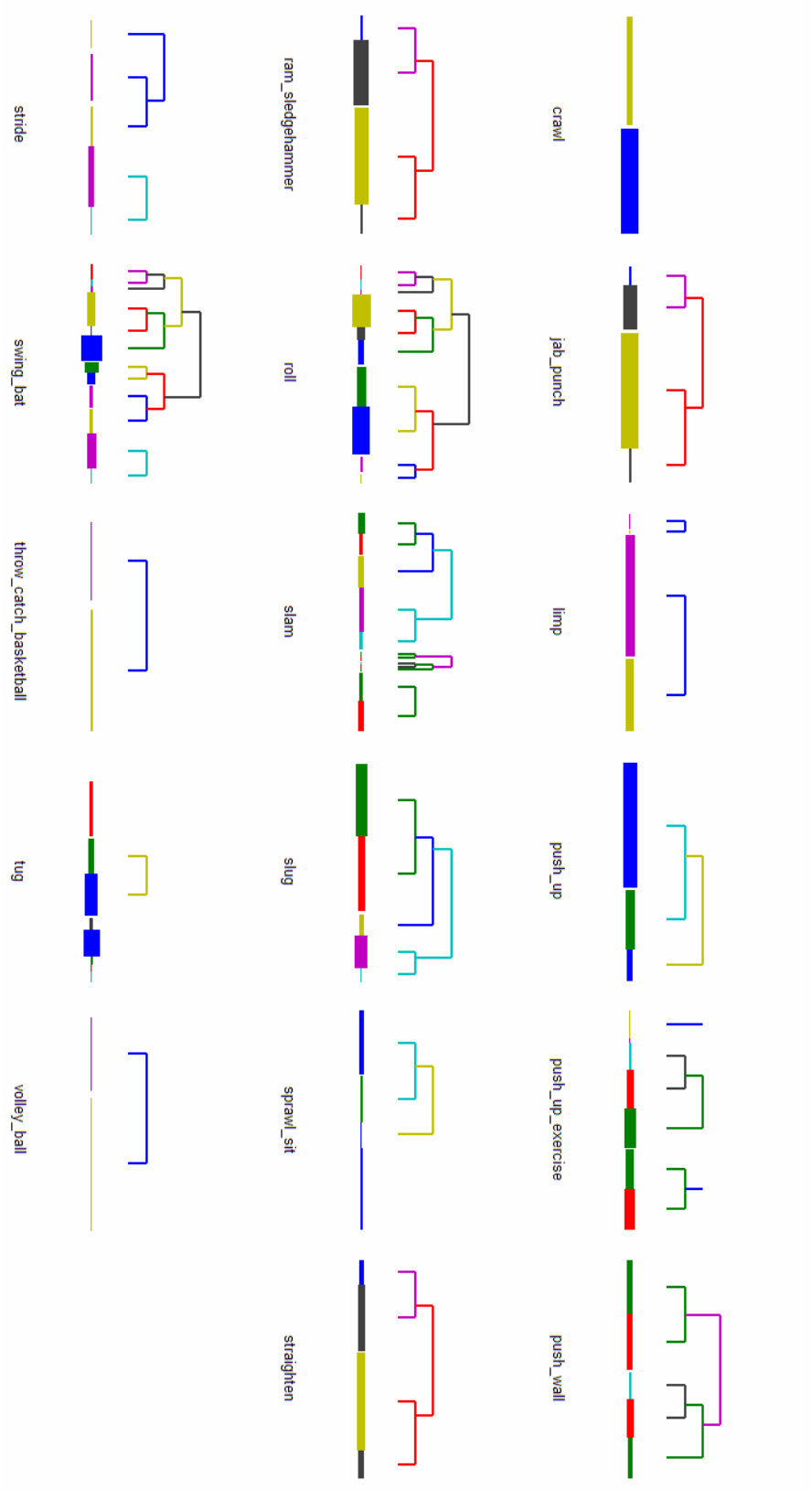
Right Shoulder X


carry_briefcase
carry_briefcase2
climb_stair
cock
collar
conk
crawl
crawl2
crouch
curtsy
curtsy2
cut
descend_stair
dig
duck
fluff_up
hew
hold_ears
hook_ball
jab
jab_punch
jam_hand
jog
jump_forward
jump_inplace
jump_sideways
kill_ball
kiss
knuckle_cir
limp2
march2
oscillate
pick_up_ball
pick_up_ball2
pick_up_object
pull
pull_back
push
push_pull
push_up
push_up_exercise
push_wall
ram_sledgehammer
rub_cir
rub_ver
run
run_backwards
run_sideways_crossing
scrape
shuffle
shuffle2
slam
slug
spread_eagle
sprint
squat
stab
stoop
straighten
stride
stride2
swagger
swing_bat
tip
tiptoe
toe_inwards
toe_inwards2
toe_outwards
toe_outwards2
trot
tug
volley_ball
walk
walk2
walk_backwards2
whack
whip

Left Elbow Z



Right Elbow X


crawl
jab_punch
limp
push_up
push_up_exercise
push_wall
ram_sledgehammer
roll
slam
slug
sprawl_sit
straighten
stride
swing_bat
throw_catch_basketball
tug
volley_ball

Left Wrist Z

Bibliography

[Ahmed at al., 2002] Ahmed, A., Hilton, A., and Mokhtarian, F. 2002, "Adaptive compression of human animation data", *EuroGraphics Conference*, Saarbrücken, Germany.

[Alon at al., 2003] Alon, J., Sclaroff, S., Kollios, G., and Pavlovic, V. 2003, "Discovering clusters in motion time-series data", *IEEE International Conference on Computer Vision and Pattern Recognition*, Madison, WI, vol. I, pp. 375-381.

[Arbib, 1992] Arbib, M. 1992, "Schema theory" in *The Encyclopedia of Artificial Intelligence*, ed. S. Shapiro, Wiley Interscience, New York, vol. 2, pp. 1427-1443.

[Arikan and Forsythe, 2002] Arikan, O. and Forsythe, D. 2002, "Interactive motion generation from examples", *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 483-490.

[Armstrong at al., 1995] Armstrong, D., Stokoe, W., and Wilcox, S. 1995, *Gesture and the Nature of Language*, Cambridge University Press, New York.

[Ashraf and Wong, 2000] Ashraf, G. and Wong, K. 2000, "Generating consistent motion transition via decoupled framespace interpolation", *Computer Graphics Forum*, vol. 19, no. 3, pp. 447-456.

[Assa at al., 2005] Assa, J., Caspi, Y., and Cohen-Or, D. 2005, "Action synopsis: Pose selection and illustration", *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 667-676.

[Atkeson and Schaal, 1997] Atkeson, C. and Schaal, S. 1997, "Robot learning from demonstration", *International Conference on Machine Learning*, Nashville, TN, pp. 12-20.

[Averbeck at al., 2003a] Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. 2003, "Neural activity in prefrontal cortex during copying geometrical shapes - I. Single cells encode shape, sequence, and metric parameters", *Experimental Brain Research*, vol. 150, no. 2, pp. 127-141.

[Averbeck at al., 2003b] Averbeck, B., Crowe, D., Chafee, M., and Georgopoulos, A. 2003, "Neural activity in prefrontal cortex during copying geometrical shapes - II. Decoding shape segments from neural ensembles", *Experimental Brain Research*, vol. 150, no. 2, pp. 142-153.

[Bailey at al., 1998] Bailey, D., Chang, N., Feldman, J., and Narayanan, S. 1998, "Extending embodied lexical development", *Annual Meeting of the Cognitive Science Society*, Madison, WI.

[Barbič at al., 2004] Barbič, J., Safonova, A., Pan, J.-Y., Faloutsos, C., Hodgins, J., and Pollard, N. 2004, "Segmenting motion capture data into distinct behaviors", *Conference on Graphics Interface*, London, Canada, pp. 185-194.

[Billard and Matarić, 2001] Billard, A. and Matarić, M. 2001, "Learning human arm movements by imitation: Evaluation of a biologically-inspired connectionist architecture", *Robotics and Autonomous Systems*, vol. 41, no. 9, pp. 1-16.

[Birdwhistell, 1970] Birdwhistell, R. 1970, *Kinesics and Context*, University of Pennsylvania Press, Philadelphia.

[Boyd and Little, 1997] Boyd, J. and Little, J. 1997, "Global versus structured interpretation of motion: Moving light displays", *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, San Juan, Puerto Rico, pp. 18-25.

[Brand and Hertzmann, 2000] Brand, M. and Hertzmann, A. 2000, "Style machines", *International Conference on Computer Graphics and Interactive Techniques*, New Orleans, LA, pp. 183-192, 2000.

[Browman and Goldstein, 1985] Browman, C. and Goldstein, L. 1985, "Dynamic modeling of phonetic structure" in *Phonetic Linguistics*, ed. V. Fromkin, Academic Press, New York, pp. 35-53.

[Browman and Goldstein, 1990] Browman, C. and Goldstein, L. 1990, "Gestural specification using dynamically-defined articulatory structures", *Journal of Phonetics*, vol. 18, no. 3, pp. 299-320.

[Buccino at al., 2004] Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C., and Rizzolatti, G. 2004, "Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI study", *Journal of Cognitive Neuroscience*, vol. 16, no. 1, pp. 114-126.

[Caelli at al., 2001] Caelli, T., McCabe, A., and Binsted, G. 2001, "On learning the shape of complex actions", *International Workshop on Visual Form*, Capri, Italy, pp. 24-39.

[Chai and Hodgins, 2005] Chai, J. and Hodgins, J. 2005, "Performance animation from low-dimensional control signals", *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 686-696.

[Chen at al., 2005] Chen, Y., Lee, J., Parent, R., and Machiraju, R. 2005, "Markerless monocular motion capture using image features and physical constraints", *Computer Graphics International*, Stony Brook, NY, pp. 36-43.

[Chenevière and Boukir, 2004] Chenevière, F. and Boukir, S. 2004, "Deformable model based data compression for gesture recognition", *International Conference on Pattern Recognition*, Cambridge, United Kingdom, vol. 4, pp. 541-544.

[Cheung at al., 2005] Cheung, V., d'Avella, A., Tresch, M., and Bizzi, E. 2005, "Central and sensory contributions to the activation and organization of muscle synergies during natural motor behaviors", *Journal of Neuroscience*, vol. 25, no. 27, pp. 6419-6434.

[Clark, 1963] Clark, W. 1963, The Antecedents of Man, Harper and Row, New York.

[Condon and Ogston, 1967] Condon, W. and Ogston, W. 1967, "A segmentation of behavior", *Journal of Psychiatric Research*, vol. 5, no. 3, pp. 221-235.

[Clynes, 1970] Clynes, M. 1970, "On being in order", *Zygon: Journal of Religion and Science*, vol. 5, no. 1, pp. 63-84.

[Cox at al., 2000] Cox, I., Miller, M., Minka, T., Papathomas, T., and Yianilos, P. 2000, "The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments", *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 20-37.

[Csuhaj-Varjú and Dassow, 1990] Csuhaj-Varjú, E. and Dassow, J. 1990, "On cooperating/distributed grammar systems", *Journal of Information Processing and Cybernetics*, vol. 26, no. 1-2, pp. 49-63.

[Darwin, 1872] Darwin, C. 1872, *The Expression of the Emotions in Man and Animals*, The University of Chicago Press, Chicago.

[d'Avella and Bizzi, 2005] d'Avella, A. and Bizzi, E. 2005, "Shared and specific muscle synergies in natural motor behaviors", *Proceedings of the National Academy of Sciences*, vol. 102, no. 8, pp. 3076-3081.

[d'Avella at al., 2003] d'Avella, A., Saltiel, P., and Bizzi, E. 2003, "Combinations of muscle synergies in the construction of a natural motor behavior", *Nature Neuroscience*, vol. 6, no. 3, pp. 300-308.

[Davis, 1970] Davis, M. 1970, "Movement characteristics of hospitalized psychiatric patients", *Annual Conference of the American Dance Therapy Association*, New York, NY, pp. 25-45.

[Davis, 1973] Davis, M. 1975, *Towards Understanding the Intrinsic in Body Movement*, Arno Press, New York.

[Deane, 1991] Deane, P. 1991, "Syntax and the brain: Neurological evidence for the spatialization of form hypothesis", *Cognitive Linguistics*, vol. 2, no. 4, pp. 361-367.

[Deane, 1993] Deane, P. 1993, *Grammar in Mind and Brain: Explorations in Cognitive Syntax*, Mouton Gruyter, Berlin-New York.

[Del Vecchio at al., 2003] Del Vecchio, D., Murray, R., and Perona, P. 2003, "Decomposition of human motion into dynamics-based primitives with application to drawing tasks", *Automatica*, vol. 39, no. 12, pp. 2085-2098.

[Dell, 1971] Dell, C. 1971, *A Primer for Movement Description*, Dance Notation Bureau, New York.

[Deutsch, 1952] Deutsch, F. 1952, "Analytic posturology", *Psychoanalytic Quarterly*, vol. 21, no. 2, pp. 196-214.

[Edelman, 1989] Edelman, G. 1989, *The Remembered Present: A Biological Theory of Consciousness*, Basic Books, New York.

[Edelman, 1992] Edelman, G. 1992, *Bright Air, Brilliant Fire: On the Matter of Mind*, Basic Books, New York.

[Edmondson, 1987] Edmondson, W. 1987, "Segments in signed languages: Do they exist and does it matter?", *International Symposium on Sign Language Research*, Lappeenranta, Finland, pp. 66-74.

[Eshkol, 1980] Eshkol, N. 1980, *50 Lessons by Dr. Moshe Feldenkrais*, The Movement Notation Society, Tel-Aviv.

[Etou at al., 2004] Etou, H., Okada, Y., and Niijma, K. 2004, "Feature preserving motion compression based on hierarchical curve simplification", *IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, vol. 2, pp. 1435-1438.

[Fellbaum, 1998] Fellbaum, C. 1998, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge.

[Fernau, 2001] Fernau, H. 2001, "PC grammar systems with terminal transmission", *Acta Informatica*, vol. 37, no. 7, pp. 511-540.

[Fishbach at al., 2005] Fishbach, A., Roy, S., Bastianen, C., Miller, L., and Houk, J. 2005, "Kinematic properties of on-line error corrections in the monkey", *Experimental Brain Research*, vol. 164, no. 4, pp. 442-457.

[Flash and Hochner, 2005] Flash, T. and Hochner, B. 2005, "Motor primitives in vertebrates and invertebrates", *Current Opinion in Neurobiology*, vol. 15, no. 6, pp. 660–666.

[Fod at al., 2002] Fod, A., Matarić, M., and Jenkins, O. 2002, "Automated derivation of primitives for movement classification", *Autonomous Robots*, vol. 12, no. 1, pp. 39-54.

[Gallese at al., 1996] Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. 1996, "Action recognition in the premotor cortex", *Brain*, vol. 119, no. 2, pp. 593-609.

[Ghez at al., 1997] Ghez, C., Favilla, M., Ghilardi, M., Gordon, J., Bermejo, R., and Pullman, S. 1997, "Discrete and continuous planning of hand movements and isometric force trajectories", *Experimental Brain Research*, vol. 115, no. 2, pp. 217-233.

[Glenberg and Kaschak, 2002] Glenberg, A. and Kaschak, M. 2002, "Grounding language in action", *Psychonomic Bulletin & Review*, vol. 9, no. 3, pp. 558-565.

[Gold, 1967] Gold, E. 1967, "Language identification in the limit", *Information and Control*, vol. 10, no. 5, pp. 447-474.

[Graziano at al., 2002] Graziano, M., Taylor, C., Moore, T., and Cooke, D. 2002, "The cortical control of movement revisited", *Neuron*, vol. 36, no. 3, pp. 349-362.

[Graziano at al., 2004] Graziano, M., Patel, K., and Taylor, C. 2004, "Mapping from motor cortex to biceps and triceps altered by elbow angle", *Journal of Neurophysiology*, vol. 92, no. 1, pp. 395-407.

[Greenfield, 1991] Greenfield, P. 1991, "Language, tools and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior", *Behavioral and Brain Sciences*, vol. 14, no. 4, pp. 531-595.

[Grinyagin at al., 2005] Grinyagin, I., Biryukova, E., and Maier, M. 2005, "Kinematic and dynamic synergies of human precision-grip movements," *Journal of Neurophysiology*, vol. 94, no. 4, pp. 2284-2294.

[Guerra-Filho, 2005] Guerra-Filho, G. 2005, "Optical motion capture: Theory and implementation", *Revista de Informática Teórica e Aplicada*, vol. 12, no. 2, pp. 61-89.

[Guerra Filho and Aloimonos, 2006a] Guerra-Filho, G. and Aloimonos, Y. 2006, "Towards a sensorimotor WordNet^{SM}: Closing the semantic gap", *International WordNet Conference*, Jeju Island, Korea.

[Guerra-Filho and Aloimonos, 2006b] Guerra-Filho, G. and Aloimonos, Y. 2006, "Understanding visuo-motor primitives for motion synthesis and analysis", *Computer Animation and Virtual Worlds*, vol. 17, no. 3-4, pp. 207-217.

[Harnard, 1990] Harnard, S. 1990, "The symbol grounding problem", *Physica D*, vol. 42, pp. 335-346.

[Harrow, 1972] Harrow, A. 1972, *A Taxonomy of the Psychomotor Domain*, David McKay Company, New York.

[Hart and Giszter, 2004] Hart, C. and Giszter, S. 2004, "Modular premotor drives and unit bursts as primitives for frog motor behaviors", *Journal of Neuroscience*, vol. 24, no. 22, pp. 5269-5282.

[Heck at al., 2006] Heck, R., Kovar, L., and Gleicher, M. 2006, "Splicing upper-body actions with locomotion", *Computer Graphics Forum*, vol. 25, no. 3, pp. 459-466.

[Hockett, 1978] Hockett, C. 1978, "In search of Jove's brow", *American Speech*, vol. 53, no. 4, pp. 243-313.

[Hopcroft and Ullman, 1979] Hopcroft, J. and Ullman, J. 1979, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Boston.

[Hoyle, 1983] Hoyle, G. 1983, *Muscles and their Neural Control*, John Wiley, New York.

[Huang at al., 2001] Huang, Q., Yokoi, K., Kajita, S., Kaneko, K., Arai, H., Koyachi, N., and Tanie, K. 2001, "Planning walking patterns for a biped robot", *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 280-289.

[Hutchinson, 1977] Hutchinson, A. 1977, *Labanotation*, Theatre Arts Books, New York.

[Ikemoto and Forsyth, 2004] Ikemoto, L. and Forsyth, D. 2004, "Enriching a motion collection by transplanting limbs", *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Grenoble, France, pp. 99-108.

[Ilg at al., 2004] Ilg, W., Bakir, G., Mezger, J., and Giese, M. 2004, "On the representation, learning and transfer of spatio-temporal movement characteristics", *International Journal of Humanoid Robotics*, vol. 1, no. 4, pp. 613-636.

[Inamura at al., 2002] Inamura, T., Toshima, I., and Nakamura, Y. 2002, "Acquiring motion elements for bidirectional computation of motion recognition and generation" in *Experimental Robotics VIII*, eds. B. Siciliano and P. Dario, Springer, Berlin, pp. 372-381.

[Ivanenko at al., 2005] Ivanenko, Y., Cappellini, G., Dominici, N., Poppele, R., and Lacquaniti, F. 2005, "Coordination of locomotion with voluntary movements in humans", *Journal of Neuroscience*, vol. 25, no. 31, pp. 7238-7253.

[Ivanov and Bobick, 2000] Ivanov, Y. and Bobick, A. 2000, "Recognition of visual activities and interactions by stochastic parsing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 852-872.

[Jackendoff, 1997] Jackendoff, R. 1997, *The Architecture of the Language Faculty*, MIT Press, Cambridge.

[Jeannerod, 1994] Jeannerod, M. 1994, "Object oriented action" in *Insights into the Reach to Grasp Movement*, eds. K. Bennett and U. Castiello, Elsevier, Amsterdam, pp. 3-15.

[Jeannerod at al., 1995] Jeannerod, M., Arbib, M., Rizzolatti, G., and Sakata, H. 1995, "Grasping objects - The cortical mechanisms of visuomotor transformation", *Trends in Neurosciences*, vol. 18, no. 7, pp. 314-320.

[Jenkins and Matarić, 2003] Jenkins, O. and Matarić, M. 2003, "Automated derivation of behavior vocabularies for autonomous humanoid motion", *International Conference on Autonomous Agents*, Melbourne, Australia, pp. 225-232.

[Jerde and Flanders, 2003] Jerde, T. and Flanders, M. 2003, "Coarticulation in fluent fingerspelling", *Journal of Neuroscience*, vol. 23, no. 6, pp. 2383-2393.

[Johnson, 1987] Johnson, M. 1987, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*, University of Chicago Press, Chicago.

[Kahol at al., 2004] Kahol, K., Tripathi, P., and Panchanathan, S. 2004, "Automated gesture segmentation from dance sequences", *IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, pp. 883-888.

[Kang at al., 2004] Kang, N., Shinohara, M., Zatsiorsky, V., and Latash, M. 2004, "Learning multi-finger synergies: An uncontrolled manifold analysis", *Experimental Brain Research*, vol. 157, no. 3, pp. 336-350.

[Kelso at al., 1986] Kelso, J., Saltzman, E., and Tuller, B. 1986, "The dynamical perspective on speech production: Data and theory", *Journal of Phonetics*, vol. 14, pp. 29-59.

[Kestenberg at al., 1971] Kestenberg, J., Marcus, H., Robbins, E., Berlowe, J., and Buelte, A. 1971, "Development of the young child as expressed through bodily movement", *Journal of the American Psychoanalytic Association*, vol. 19, no. 4, pp. 746-764.

[Kien, 1992] Kien, J. 1992, "Temporal segmentation in the motor system, symbolization, and the evolution of language", *Annual Meeting of the Language Origin Society*, Cambridge, United Kingdom.

[Kimura, 1981] Kimura, D. 1981, "Neural mechanisms in manual signing", *Sign Language Studies*, vol. 33, pp. 291-312.

[Ko and Badler, 1996] Ko, H. and Badler, N. 1996, "Animating human locomotion with inverse dynamics", *IEEE Computer Graphics and Applications*, vol. 16, no. 2, pp. 50-59.

[Kovar and Gleicher, 2004] Kovar, L. and Gleicher, M. 2004, "Automated extraction and parameterization of motions in large data sets", *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 559-568.

[Kovar at al., 2002] Kovar, L., Gleicher, M., and Pighin, F. 2002, "Motion graphs", *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 473-482.

[Kuniyoshi at al., 1994] Kuniyoshi, Y., Inaba, M., and Inoue, H. 1994, "Learning by watching: Extracting reusable task knowledge from visual observation of human performance", *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 799-822.

[Lakoff, 1987] Lakoff, G. 1987, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago.

[Langacker, 1991] Langacker, R. 1991, *Foundations of Cognitive Grammar*, Volume II, Stanford University Press, Stanford.

[Latecki and Lakämper, 1999] Latecki, L. and Lakämper, R. 1999, "Convexity rule for shape decomposition based on discrete contour evolution", *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 441-454.

[Lee at al., 2002] Lee, J., Chai, J., Reitsma, P., Hodgins, J., and Pollard, N. 2002, "Interactive control of avatars animated with human motion data", *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 491-500.

[Liddell, 1984] Liddell, S. 1984, "Think and believe: Sequentiality in American Sign Language", *Language*, vol. 60, no. 2, pp. 372-399.

[Liddell and Johnson, 1989] Liddell, S. and Johnson, R. 1989, "American Sign Language: The phonological base", *Sign Language Studies*, vol. 64, pp. 195-277.

[Lim and Thalmann, 2001] Lim, I. and Thalmann, D. 2001, "Key-posture extraction out of human motion data by curve simplification", *International Conference of the IEEE Engineering in Medicine and Biology Society*, Istanbul, Turkey, vol. 2, pp. 1167-1169.

[Mahl, 1968] Mahl, G. 1968, "Gestures and body movements in interviews" in *Research in Psychotherapy*, ed. J. Shlien, American Psychological Association, Washington D.C., vol. 3, pp. 295-346.

[Matarić, 2002] Matarić, M. 2002, "Visuo-motor primitives as a basis for learning by imitation: Linking perception to action and biology to robotics" in *Imitation in Animals and Artifacts*, eds. K. Dautenhahn and C. Nehaniv, MIT Press, Cambridge, pp. 392-422.

[Matsui at al., 2005] Matsui, D., Minato, T., MacDorman, K., and Ishiguro, H. 2005, "Generating natural motion in an android by mapping human motion", *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Canada, pp. 3301-3308.

[McGee, 1978] McGee, V. 1978, "Multidimensional scaling of *n* sets of similarity measures: A nonmetric individual differences approach", *Multivariate Behavioral Research*, vol. 3, pp. 233-248.

[McNeill, 1985] McNeill, D. 1985, "So you think gestures are nonverbal", *Psychological Review*, vol. 92, no. 3, pp. 350-371.

[Meersman and Rozenberg, 1978] Meersman, R. and Rozenberg, G. 1978, "Cooperating grammar systems", *Lecture Notes in Computer Science*, vol. 64, pp. 364-374.

[Menache, 2000] Menache, A. 2000, *Understanding Motion Capture for Computer Animation and Video Games*, Morgan Kaufmann, San Francisco.

[Mezger at al., 2005] Mezger, J., Ilg, W., and Giese, M. 2005, "Trajectory synthesis by hierarchical spatio-temporal correspondence: Comparison of different methods",

ACM Symposium on Applied Perception in Graphics and Visualization, A Coruña, Spain, pp. 25-32.

[Mörchen at al., 2005] Mörchen, F., Ultsch, A., and Hoos, O. 2005, "Extracting interpretable muscle activation patterns with time series knowledge mining", *International Journal of Knowledge-Base and Intelligent Engineering Systems*, vol. 9, no. 3, pp. 197-208.

[Mori and Uehara, 2001] Mori, T. and Uehara, K. 2001, "Extraction of primitive motion and discovery of association rules from motion data", *IEEE International Workshop on Robot and Human Interactive Communication*, Bordeaux and Paris, France, pp. 200-206.

[Mowrey and Pagliuca, 1995] Mowrey, R. and Pagliuca, W. 1995, "The reductive character of articulatory evolution", *Rivista di Linguistica*, vol. 7, no. 1, pp. 37-124.

[Mussa-Ivaldi and Bizzi, 2000] Mussa-Ivaldi, F. and Bizzi, E. 2000, "Motor learning through the combination of primitives", *Philosophical Transactions of the Royal Society of London B - Biological Sciences*, vol. 355, no. 1404, pp. 1755-1769.

[Mussa-Ivaldi and Solla, 2004] Mussa-Ivaldi, F. and Solla, S. 2004, "Neural primitives for motion control", *IEEE Journal of Oceanic Engineering*, vol. 29, no. 3, pp. 640-650.

[Naka at al., 1999] Naka, T., Mochizuki, Y., Hijiri, T., Cornish, T., and Asahara, S. 1999, "A compression/decompression method for streaming based humanoid animation", *Symposium on Virtual Reality Modeling Language*, Paderborn, Germany, pp. 63-70.

[Nakazawa at al., 2002] Nakazawa, A., Nakaoka, S., Ikeuchi, K., Yokoi, K. 2002, "Imitating human dance motions through motion structure analysis", IEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne, Switzerland, pp. 2539-2544.

[Nevill-Manning and Witten, 1997] Nevill-Manning, C. and Witten, I. 1997, "Identifying hierarchical structure in sequences: A linear-time algorithm", *Journal of Artificial Intelligence Research*, vol. 7, pp. 67-82.

[Nishitani at al., 2005] Nishitani, N., Schürmann, M., Amunts, K., and Hari, R. 2005, "Broca's region: From action to language", *Physiology*, vol. 20, no. 1, pp. 60-69.

[North, 1971] North, M. 1971, *Personality Assessment through Movement*, Mac-Donald & Evans, London.

[Parekh and Honavar, 2000] Parekh, R. and Honavar, V. 2000, "Grammar inference, automata induction, and language acquisition" in *The Handbook of Natural Language Processing*, eds. R. Dale, H. Moisl, and H. Somers, Marcel Dekker, New York, pp. 727-764.

[Pasalar at al., 2005] Pasalar, S., Roitman, A., and Ebner, T. 2005, "Effects of speeds and force fields on submovements during circular manual tracking in humans", *Experimental Brain Research*, vol. 163, no. 2, pp. 214-225.

[Păun, 1993] Păun, G. 1993, "On the synchronization in parallel communicating grammar systems", *Acta Informatica*, vol. 30, no. 4, pp. 351-367.

[Păun and Sântean, 1989] Păun, G. and Sântean, L. 1989, "Parallel communicating grammar systems: The regular case", *Annals of the University of Bucharest, Mathematics-Informatics Series*, vol. 38, no. 2, pp. 55-63.

[Perlin, 1995] Perlin, K. 1995, "Real time responsive animation with personality", *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 1, pp. 5-15.

[Perlmutter, 1988] Perlmutter, D. 1988, "A mosaic theory of American Sign Language syllable structure", *Conference on Theoretical Issues in Sign Language Research*, Gallaudet University, Washington D.C..

[Perrett at al., 1989] Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hietanen, J., and Ortega, J. 1989, "Frameworks of analysis for the neural representation of animate objects and actions", *Journal of Experimental Biology*, vol. 146, no. 1, pp. 87-113.

[Petitto and Marentette, 1991] Petitto, L. and Marentette, P. 1991, "Babbling in the manual mode: Evidence for the ontogeny of language", *Science*, vol. 251, no. 5000, pp. 1493-1496.

[Poizner at al., 1987] Poizner, H., Klima, E., and Bellugi, U. 1987, *What the Hands Reveal about the Brain*, MIT Press, Cambridge.

[Pollard at al., 2002] Pollard, N., Hodgins, J., Riley, M., and Atkeson, C. 2002, "Adapting human motion for the control of a humanoid robot", *International Conference on Robotics and Automation*, Washington, D.C., vol. 2, pp. 1390-1397.

[Rao and Shah, 2001] Rao, C. and Shah, M. 2001, "View-invariance in action recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai Island, Hawaii, vol. 2, pp. 316-321.

[Reich, 1949] Reich, W. 1949, *Character Analysis*, Farrar, Straus & Giroux (The Noonday Press), New York.

[Rohrer at al., 2002] Rohrer, B., Fasoli, S., Krebs, H., Hugh, R., Volpe, B., Frontera, W., Stein, J., and Hogan, N. 2002, "Movement smoothness changes during stroke recovery", *Journal of Neuroscience*, vol. 22, no. 18, pp. 8297-8304.

[Roitman at al., 2004] Roitman A., Massaquoi, S., Takahashi, K., and Ebner, T. 2004, "Kinematic analysis of manual tracking in monkeys: Characterization of movement intermittencies during a circular tracking task", *Journal of Neurophysiology*, vol. 91, no. 2, pp. 901-911.

[Rose at al., 1998] Rose, C., Cohen, M., and Bodenheimer, B. 1998, "Verbs and adverbs: Mutidimensional motion interpolation", *IEEE Computer Graphics and Applications*, vol. 18, no. 5, pp. 32-40.

[Rose at al., 1996] Rose, C., Guenter, B., Bodenheimer, B., and Cohen, M. 1996, "Efficient generation of motion transitions using spacetime constraints", *International Conference on Computer Graphics and Interactive Techniques*, New York, NY, pp. 147-154.

[Rui at al., 1998] Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. 1998, "Relevance feedback: A power tool for interactive content-based image retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644-655.

[Samejima at al., 2002] Samejima, K., Katagiri, K., Doya, K., and Kawato, M. 2002, "Symbolization and imitation learning of motion sequence using competitive modules", *The Transactions of the Institute of Electronics, Information and Communication Engineers*, vol. J85-D-II, no. 1, pp. 90-100.

[Sandler, 1986] Sandler, W. 1986, "The spreading hand autosegment of American Sign Language", *Sign Language Studies*, vol. 50, pp. 1-28.

[Saux, 1999] Saux, E. 1999, "Data reduction of polygonal curves using B-splines", *Computer-Aided Design*, vol. 31, no. 8, pp. 507-515.

[Schaal, 1999] Schaal, S. 1999, "Is imitation learning a route to humanoid robots?", *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 233-242.

[Schaal at al., 2003] Schaal, S., Ijspeert, and Billard, A. 2003, "Computational approaches to motor learning by imitation", *Philosophical Transactions of the Royal Society of London B - Biological Sciences*, vol. 358, no. 1431, pp. 537-547.

[Sidenbladh at al., 2002] Sidenbladh, H., Black, M., and Sigal, M. 2002, "Implicit probabilistic models of human motion for synthesis and tracking", *European Conference on Computer Vision*, Copenhagen, Denmark, pp. 784-800.

[Siskind, 2001] Siskind, J. 2001, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic", *Journal of Artificial Intelligence Research*, vol. 15, pp. 31-90.

[Solan at al., 2005] Solan, Z., Horn, D., Ruppin, E., and Edelman, S. 2005, "Unsupervised learning of natural languages", *Proceedings of the National Academy of Sciences*, vol. 102, no. 33, pp. 11629-11634.

[Sosík and Štýbnar, 1997] Sosík, P. and Štýbnar, L. 1997, "Grammatical inference of colonies", *Lecture Notes in Computer Science*, vol. 1218, pp. 236-246.

[Stein, 2005] Stein, P. 2005, "Neuronal control of turtle hindlimb motor rhythms", *Journal of Comparative Physiology A*, vol. 191, no. 3, pp. 213-229.

[Stokoe, 1960] Stokoe, W. 1960, "Sign language structure: An outline of the visual communication systems of the American deaf", *Studies in Linguistics: Occasional Papers*, no. 8, Dept. of Anthropology and Linguistics, University of Buffalo, Buffalo.

[Stokoe at al., 1965] Stokoe, W., Croneberg, C., and Casterline, D. 1965, *A Dictionary of American Sign Language on Linguistic Principles*, Linstok Press, Silver Spring.

[Stuart and Bradley, 1998] Stuart, J. and Bradley, E. 1998, "Learning the grammar of dance", *International Conference on Machine Learning*, Madison, WI, pp. 547-555.

[Studdert-Kennedy, 1985] Studdert-Kennedy, M. 1985, "Perceiving phonetic events" in *Perspectives and Change*, eds. W. Warren and R. Shaw, Lawrence Erlbaum, Hillsdale, pp. 139-156.

[Studdert-Kennedy, 1987] Studdert-Kennedy, M. 1987, "The phoneme as a perceptuomotor structure" in *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing*, ed. D. Allport, Academic Press, London, pp. 67-84.

[Sudarsky and House, 1998] Sudarsky, S. and House, D. 1998, "Motion capture data manipulation and reuse via B-splines", *Lecture Notes in Artificial Intelligence*, vol. 1537, pp. 55-69.

[Ting and Macpherson, 2005] Ting, L. and Macpherson, J. 2005, "A limited set of muscle synergies for force control during a postural task", *Journal of Neurophysiology*, vol. 93, no. 1, pp. 609-613.

[Togawa and Okuda, 2005] Togawa, H. and Okuda, M. 2005, "Position-based keyframe selection for human motion animation", *International Conference on Parallel and Distributed Systems*, Fukuoka, Japan, vol. 2, pp. 182-185.

[Tresch at al., 1999] Tresch, M., Saltiel, P., and Bizzi, E. 1999, "The construction of movement by the spinal cord", *Nature Neuroscience*, vol. 2, no. 2, pp. 162-167.

[Ude at al., 2000] Ude, A., Man, C., Riley, M., and Atkeson, C. 2000, "Automatic generation of kinematic models for the conversion of human motion capture data into humanoid robot motion", *IEEE-RAS International Conference on Humanoid Robots*, Boston, MA, pp. 2223-2228.

[Varela at al., 1991] Varela, F., Thompson, E., and Rosch, E. 1991, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge.

[Viviani, 1986] Viviani, P. 1986, "Do units of motor action really exist?" in *Generation and Modulation of Action Patterns*, eds. H. Heuer and C. Fromm, Springer, New York, pp. 201-216.

[Volterra and Erting, 1990] Volterra, V. and Erting, C. 1990, *From Gesture to Language in Hearing and Deaf Children*, Springer-Verlag, Berlin.

[Yang at al., 1997] Yang, J., Xu, Y., and Chen, C. 1997, "Human action learning via hidden Markov model", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 27, no. 1, pp. 34-44.

[Wang at al., 2001] Wang, T.-S., Shum, H.-Y., Xu, Y.-Q., and Zheng, N.-N. 2001, "Unsupervised analysis of human gestures", *IEEE Pacific Rim Conference on Multimedia*, Beijing, China, pp. 174-181.

[Wilbur, 1987] Wilbur, R. 1987, *American Sign Language: Linguistic and Applied Dimensions*, College Hill Press, Boston.

[Wiley and Hahn, 1997] Wiley, D. and Hahn, J. 1997, "Interpolation synthesis of articulated figure motion", *IEEE Computer Graphics and Applications*, vol. 17, no. 6, pp. 39-45.

[Wolff, 1945] Wolff, C. 1945, *A Psychology of Gesture*, Methuen & Co., London.

[Wolff, 1988] Wolff, J. 1988, "Learning syntax and meanings through optimization and distributional analysis" in *Categories and Processes in Language Acquisition*, eds. Y. Levy, I. Schlesinger, and M. Braine, Lawrence Erlbaum, Hillsdale, pp. 179-215.

[Zhang and Chen, 2002] Zhang, C. and Chen, T. 2002, "Active learning framework for content-based information retrieval", *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 260-268.