A Language for Human Action

Gutemberg Guerra-Filho **Yiannis Aloimonos**
Computer Vision Laboratory
Department of Computer Science
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
guerra@cs.umd.edu, yiannis@cfar.umd.edu

Abstract

Human-centered computing (HCC) is centered on humans and what they do, i.e. human actions. Thus, developing an infrastructure for HCC requires understanding human action, at some level of detail. We need to be able to talk about actions, synthesize actions, recognize actions, manipulate actions, imitate actions, imagine and predict actions. How could we achieve this in a principled fashion?

This paper proposes that the space of human actions has a linguistic structure. This is a sensory-motor space consisting of the evolution of the 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 visual and motor information. In **phonology**, we define atomic segments (kinetemes) that are used to compose human activity. In **morphology**, we propose parallel learning to incorporate associative learning into a language inference approach. Parallel learning solves the problem of overgeneralization and is effective in identifying the active joints and motion patterns in a particular action. In **syntax**, we point out some of the basic constraints for sentence formation. Finally, we demonstrate this linguistic framework on a praxicon of 200 human actions (motion capture data obtained by a suit) and we discuss the implications of HAL on HCC.

1 Introduction: Why Human Action?

Human-Centered Computing (HCC) 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.

Another fundamental component of anthropocentric systems is the consideration of human sensory-motor skills in a wide range of activities. This way, the interface between artificial agents and human users accounts for perception and motion in a novel interaction paradigm. This paradigm leads to behavior understanding through representations (cognitive models) which allow content description and, ultimately, the integration of real and virtual worlds.

Perhaps one of the most important aspects of HCC is the need for computers to be able to share with humans a conceptual system. Concepts are the elementary units of reason and linguistic meaning. A commonly held philosophical position is 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 cognitive hidden mechanisms and knowledge incorporation. There exists a variety of studies in many disciplines (neurophysiology, psychophysics, cognitive linguistics) suggesting that indeed the human sensory-motor system is deeply involved in concept representations.

The functionality of Broca's region in the brain [Nishitani at al., 2005] and the mirror neurons theory [Gallese at al., 1996] suggests that perception and ac-

tion share the same symbolic structure that provides common ground for sensory-motor tasks (e.g. recognition and motor planning) and higher-level activities. Furthermore, spoken language and visible movement use a similar cognitive substrate based on the embodiment of grammatical processing. There is evidence that language is grounded on the motor system [Glenberg and Kaschak, 2002], which implies the possibility of a linguistic framework for a grounded representation.

In a nutshell, the computers in a HCC environment could become powerful if they possess models of human actions. In this paper, 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 is at 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 is able to learn skills through imitation, better interact with humans, and understand human activities. This understanding includes reasoning and the association of meaning to concrete concepts. The closing of this semantic gap involves the grounding of concepts on the sensory-motor information. The grounding process may start from video, where objects are detected and recognized. At this level, human body parts are features extracted from visual input and, consequently, human movement is captured. In this paper, we are interested in human actions corresponding to general observable voluntary movement.

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 toward visual and motor spaces. Our thesis is that there exists a language (in a formal sense) that describes all human action. We show how we could obtain this language using empirical data. The phonology of human movement involves the segmentation problem, the symbolization problem, and an evaluation system.

In our linguistic framework, we aim initially to find movement primitives as basic atoms. Fod at al. [2002] find primitives by k-means clustering the projection of high-dimensional segment vectors onto a reduced subspace. Kahol at al. [2004] use the local minimum in total body force to detect segment boundaries. In Nakazawa at al. [2002], similarities of motion segments are measured according to a dynamic programming distance and clustered with a nearest-neighbor algorithm. Wang at al. [2001] segment gestures with the local minima of velocity and local maxima of change in direction. The segments are hierarchically clustered into classes using Hidden Markov Models to compute a metric. A lexicon is inferred from the resulting discrete symbol sequence through a language learning approach.

The morphology of human activity is posed here as a grammatical inference problem. Language learning consists of grammar induction and structure generalization. Current approaches [Nevill-Manning and Witten, 1997; Solan at al., 2005; Wolff, 1988] account only for sequential learning. In this paper, we introduce parallel learning.

The experimental validation of our linguistic framework is performed in a motion capture database. Our motion capture database contains around 200 different actions corresponding to verbs associated with voluntary observable movement. The actions are not limited to any specific domain. Instead, the database includes actions of several types: manipulative (prehension and dexterity), non-locomotor, locomotor, and interaction.

The paper follows with the concept of kinetology and its five basic properties. The morphology of human movement is described in section 3 through parallel language learning. In section 4, we discuss the syntax of human activity. Section 5 summarizes our main results and indicates future research.

2 Kinetology

Additionally to a geometric representation for human movement, a kinetological system consists of segmentation, symbolization, and principles.

2.1 Segmentation

Automatic segmentation is the decomposition of action sequences into movement primitives. Theses primitives are meaningful atomic elements with characteristic properties which stay constant within a segment.

In order to segment human movement, we consider each joint actuator independently. Characteristic data for an action is shown in Fig. 1. For each joint (vertical axis), there are at most three varying rotation angles over time (horizontal axis). Each joint angle is represented as a 1D function over time (coded with color in Fig. 1).

An actuator is associated with a joint angle specifying the original 3D motion of the actuator according to a geometric representation (see Fig. 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. We define a state according to the sign of derivatives of a joint angle function (see Fig. 2b).

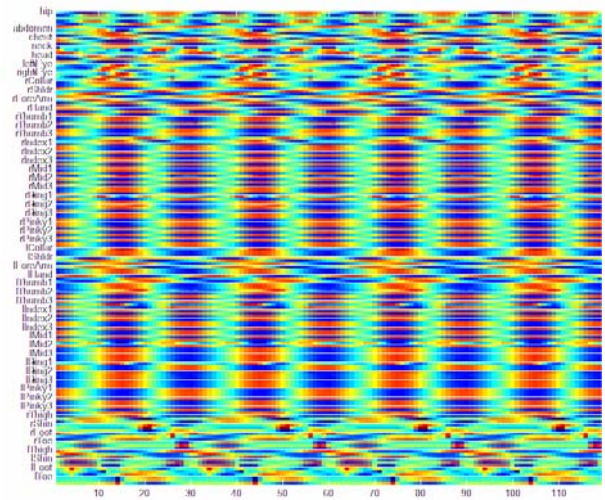
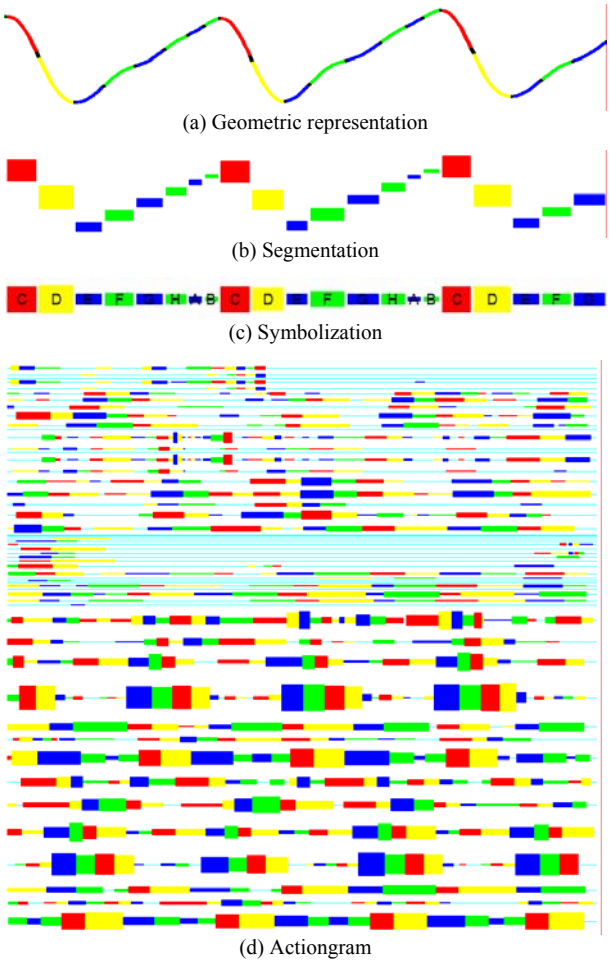


Fig. 1: Walking while playing with the fingers (finger snapping).

The derivatives used in our segmentation are velocity (first derivative) and acceleration (second derivative). This leads to a four-state system: positive velocity/positive acceleration (**B**lue), positive velocity/negative acceleration (**G**reen), negative velocity/positive acceleration (**Y**ellow), and negative velocity/negative acceleration (**R**ed). Each segment corresponds to an atom α, where $\alpha \in \{\mathbf{B}, \mathbf{G}, \mathbf{Y}, \mathbf{R}\}$ is a symbol associated with the segment's state.

The representation has a qualitative aspect, the state of each segment, and a quantitative aspect corresponding to the time length and angular displacement of each segment. The qualitative aspect is depicted with colors, while the quantitative aspect is represented by the line segment length and thickness for time length and displacement, respectively (see Fig. 2b-c).



(a) Geometric representation

(b) Segmentation

(c) Symbolization

(d) Actiongram

Fig. 2: A kinetological system.

2.2 Symbolization

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 spatio-temporal structure (see Fig. 2c). A simple way to perform symbolization is clustering using an appropriate similarity distance for segments with the same atomic state.

The symbolization results in a set of strings for the whole body motion that defines a structure denoted as *actiongram* (see Fig. 2d). An actiongram A has n strings $A_1, \ldots, A_n$. Each string A_i corresponds to an actuator and contains a (possibly different) number of m_i symbols. Each symbol $A_i(j)$ is associated with a segment, its symbol, time period, and angular displacement.

2.3 Principles

Besides sensory-motor primitives, we suggest five kinetological properties (compactness, view-invariance, reproducibility, selectivity, and reconstructivity) in order to evaluate our approach, and any other. In [Guerra-Filho and Aloimonos, 2006], we discuss these principles in detail and demonstrate that our segmentation method and primitives possess these properties.

The *compactness* principle is related to describing a human activity with the least number of atoms. Compactness is achieved through segmentation which reduces the number of parameters in the representation. Our segmentation approach was implemented as a compression method for motion data. The compression efficiency of our algorithm was tested on several different actions. The median compression rate for motion

files was 3.698% of the original file size. The best compression is achieved for actions with smooth movement. Further compression could be achieved with the use of symbolization.

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.

The view-invariance evaluation requires a 2D projected version of the initial representative function according to varying viewpoints. A circular surrounding configuration of viewpoints is used. 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 (see Fig. 3). For any joint and any action in our database, the graph demonstrates a high view-invariance measure for our segmentation process (exception only at segment's borders and two degenerated viewpoints).

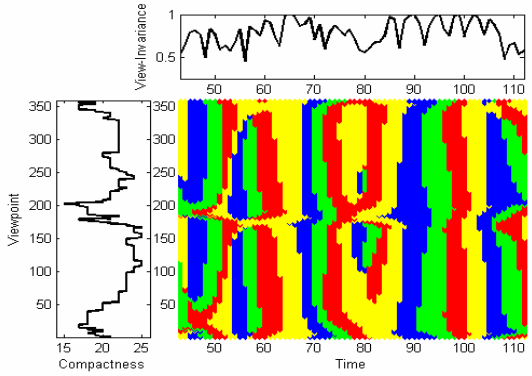


Fig. 3: View-invariance of the left knee flexion/extension angle.

Reproducibility requires an action to have the same description even when a different performance of this action is considered. A kinetological system is reproducible when the same symbolic representation is associated with the same action performed at different occasions (intra-personal) or by different subjects (inter-personal).

In order to evaluate the reproducibility of our kinetological system, we used a human gait data for 16 subjects covering males and females at several ages. This data was obtained from a normative gait database (http://physio.curtin.edu.au:16080/cga/data/index.html). 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. The reproducibility measure is very high for the joint angles which play a primary role in the walking action (see Fig. 4). The identification of the intrinsic variables of an action is a byproduct of the reproducibility requirement of a kinetological system.

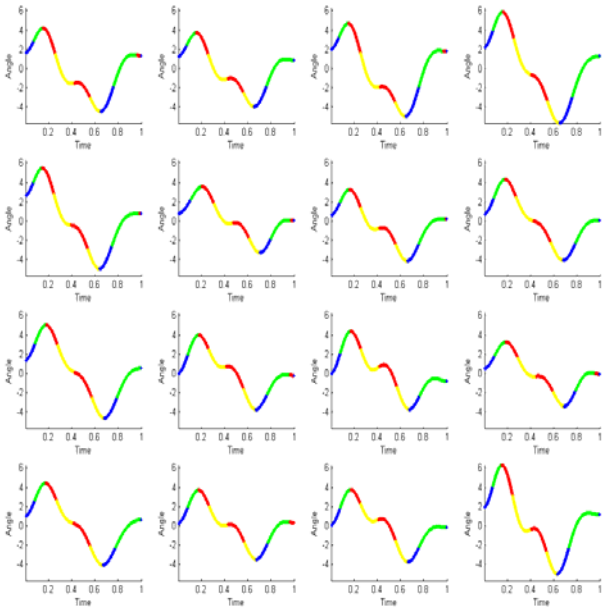


Fig. 4: Reproducibility of the pelvic obliquity during gait.

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. We compare our 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 (see Fig. 5). The different actions are clearly represented by different structures.

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

Fig. 5: Selectivity: different representations for three distinct actions.

Reconstructivity is associated with the ability to reconstruct the original movement signal up to an ap-

proximation factor from a compact representation. We propose a reconstruction method based on a novel interpolation algorithm which considers 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. Each possible sequence of three segments corresponds to two equations associated with first and second derivatives at border points of the center segment.

A simple model for the joint angle function during a segment is a polynomial. The least degree polynomial satisfying all the constraints is a fourth degree polynomial. This way, the reconstruction process needs to find five parameters. 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. These values are obtained from the time length and the angular displacement of each segment. The last free variable can be determined using some criteria such as jerk minimization.

We implemented this reconstruction scheme as a decompression method for motion data (see Fig. 6). The average error for all joints in our motion database was about 0.823°.

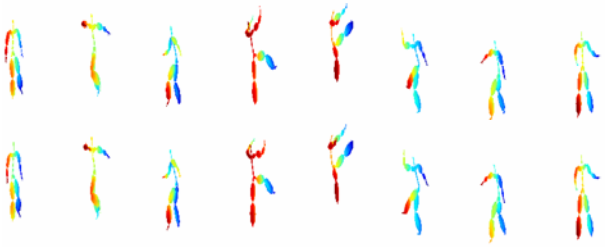


Fig. 6: Reconstructivity: original (top) and decompressed (bottom) motion sequences.

3 Morphology

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 (lexicon of human movement or praxis) 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.

The morphology of a specific human activity should include the set of joint actuators involved in the activity, the synchronization rules among these actuators, and the motion pattern associated with each actuator. In order to learn the morphology of an action, an actiongram associated with several repeated performances of this action is given as input. 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.

A *Parallel Communicating Grammar System* (PCGS) consists of several grammar components working simultaneously in a synchronized manner [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.

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 is based on a PCGS with rule synchronization [Păun, 1993]. The synchronization among rules in different components is modeled as a set of tuples of rules (one rule for each component), where rules in a tuple are derived simultaneously.

A PSGS consists in a set of Context-Free Grammars (CFGs) (see Fig. 7) related by synchronized rules. This grammar models a system with a set A of different strings A_i occurring at the same time: an actiongram. Each string A_i corresponds to the language inferred for a component grammar G_i modeling an actuator.

Fig. 7: A CFG component of a PSGS shown as a binary tree forest.

Our Parallel Learning (PAL) algorithm computes the digram frequency within each string independently. A new rule is created for the digram d with the maximum frequency. The algorithm replaces each occurrence of d in its string A_i with the created non-terminal N_c. The new non-terminal is associated with the time period corresponding to the union of the periods of both symbols in the digram.

The non-terminal N_c is checked for possible synchronized rules with non-terminals in the CFGs of other strings (see Fig. 8). Synchronization between two non-terminals (N_c and N_k) of different CFGs requires these non-terminals to have an intersecting time period 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 of their occurrences in the associated strings. Furthermore, any two mapped occurrences must correspond to intersecting time periods. The final components of the PSGS are the CFGs with synchronized rules.

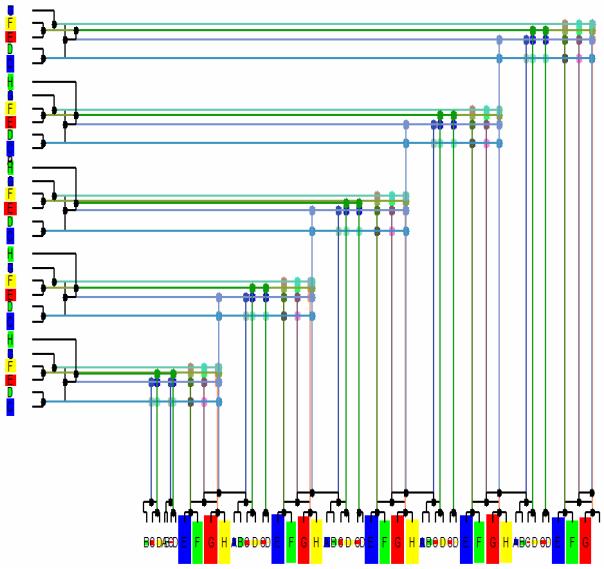
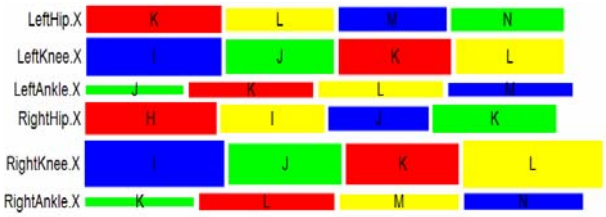


Fig. 8: Two CFGs (corresponding to hip and knee flexion/extension) related by synchronized rules of a PSGS.

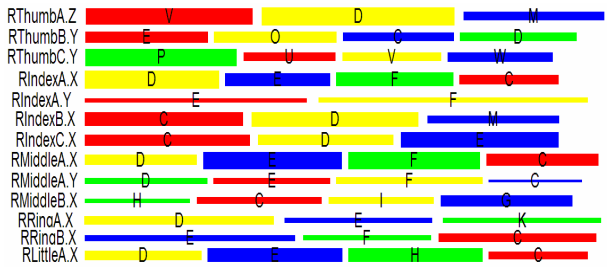
Given an actiongram of a human activity, parallel learning selects a subset of the actiongram which projects the whole action only into the intrinsic joint angles and motion patterns of the action. This process was performed in each action of our motion database and we successfully identified the morphemes in our database, i.e., the joints participating in each action, the motion patterns (kinetemes), and their synchronization with movement in other joints (see Fig. 9).

4 Syntax

The Subject-Verb-Object (SVO) pattern of syntax is a reflection of the patterns of cause and effect. 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); the action or predicate is the motion of those parts (verb).



(a) Climb stair



(b) Grab

Fig. 9: Sample morpheme examples learned for human actions.

A noun in a HAL sentence corresponds to the body parts active during the execution of a human activity and to the possible objects involved passively in the action. A noun is represented by a binary string signaling these joints (see Fig. 10). 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. The initial posture for a HAL sentence is analogous to an adjective which further describes (modifies) the active joints (nouns) in the sentence.

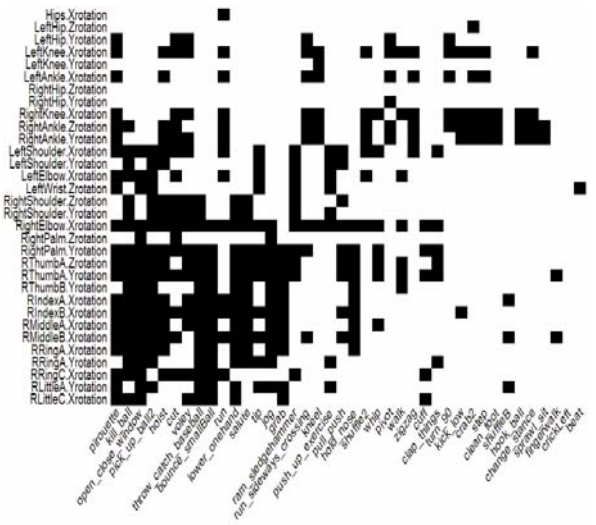


Fig. 10: HAL nouns extracted from about 200 actions in a praxicon.

The sentence verb represents the changes each active joint experiences during the action execution. The representation for a HAL verb was discussed in the previous sections. A HAL adverb models the variation in the execution of each segment in a verb. The adverb modifies the verb with the purpose of generalizing the motion. The motion of a segment is represented in a space with a reduced dimensionality, where adverbs are learned.

Given the morphology of each action in our database, we may infer additional structure on the lexical categories of movement. Further learning of the most frequent sets of joints that are active in all actions and the corresponding initial poses will lead to higher-level nouns and adjectives.

In this sense, HAL verbs for a particular joint actuator may have a common structure. Some verbs share the same kineteme (depicted as black segments in Fig. 11). This way, the morphological grammars become even more compact with a few kinetemes required to represent all motions.

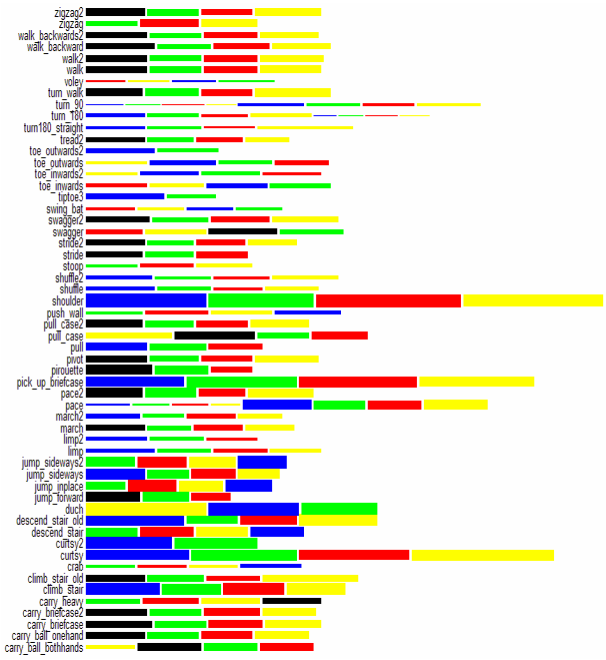


Fig. 11: HAL verbs for the left knee in our motion database.

5 Conclusion

In this paper, we advocated a linguistic framework for the representation of human activity. We introduced the concept of a kinetological system. We proposed a segmentation approach. As an evaluation method, we suggested five basic properties for a kinetological system: compactness, view-invariance, reproducibility, selectivity, and reconstructivity.

In morphology, we proposed parallel learning to incorporate associative learning with our language learning approach. Parallel learning is effective in identifying the active joints, motion patterns (kinetemes), and synchronization (coordination) in a particular action. This way, we built a praxicon with about 200 actions. The praxicon was compressed with our symbolic representations and grammatical learning was used to induce the morpheme of each action in the database.

Our intent is to provide a flexible representation, proposed here as HAL, to allow the capabilities of recognition and generation of hundreds of human actions modeled in a compact structure. This structure; organized in terms of syntax, morphology, and kinetology; has the flexibility required to handle a large number of behaviors using the parsing and generation aspects of a language. Further work amounts to the exploration of the morphological organization of a praxicon towards the discovery of more structure in the human activity language. We also expect more development concerning human movement syntax from the empirical study of this praxicon.

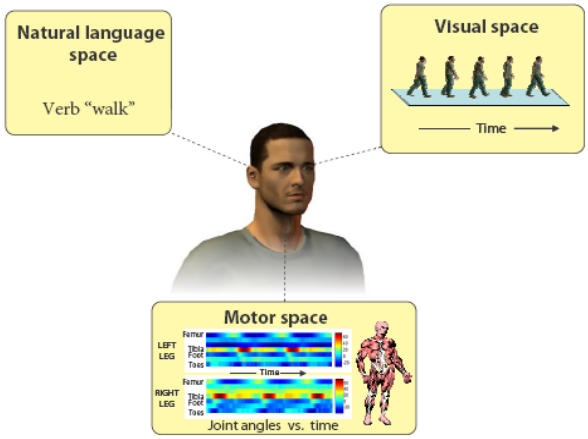


Fig. 12: Three language spaces for human action.

In conclusion, actions are represented in at least three spaces: the visual space, the motor space, and the natural language space (see Fig. 12). Therefore, we can imagine that actions possess at least three languages: a *visual* language, a *motor* language, and a *natural* language. The visual language allows us to see and understand actions, the motor language allows us to produce actions, and the natural language allows us to talk about actions. In this paper, we essentially studied a language that maps to the lower-level visual 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 translation problems. For example, (a) video annotation: creating text descriptions of activity from a video, (b) natural-language-driven character animation (computer graphics), (c) training robots by imitation using video, or (d) control of robots with natural-language. These problems are at the kernel of HCC.

References

[Fod at al., 2002] A. Fod, M. Matarić, and O. Jenkins. Automated derivation of primitives for movement classification. *Autonomous Robots*, 12(1):39-54, 2002.

[Gallese at al., 1996] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593-609, 1996.

[Glenberg and Kaschak, 2002] A. Glenberg and M. Kaschak. Grounding language in action. *Psychonomic Bulletin & Review*, 9(3):558-565, 2002.

[Guerra-Filho and Aloimonos, 2006] G. Guerra-Filho and Y. Aloimonos, Understanding visuo-motor primitives for motion synthesis and analysis. *Computer Animation and Virtual Worlds*, 17(3-4):207-217, 2006.

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

[Nakazawa at al., 2002] A. Nakazawa, S. Nakaoka, K. Ikeuchi, and K. Yokoi. Imitating human dance motions through motion structure analysis. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2539-2544, 2002.

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

[Nishitani at al., 2005] N. Nishitani, M. Schurmann, K. Amunts, and R. Hari. Broca's region: From action to language. *Physiology*, 20:60-69, 2005.

[Păun and Sântean, 1989] G. Păun and L. Sântean. Parallel communicating grammar systems: The regular case. *Annals of the University of Bucharest*, Mathematics-Informatics Series, 38(2):55-63, 1989.

[Păun, 1993] G. Păun. On the synchronization in parallel communicating grammar systems. *Acta Informatica*, 30(4):351-367, 1993.

[Solan at al., 2005] Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised learning of natural languages. *Proceedings of National Academy of Sciences*, 102(33):11629-11634, 2005.

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

[Wolff, 1988] J. Wolff. Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I. Schlesinger, and M. Braine (Eds.), Categories and Processes in Language Acquisition. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pages 179-215, 1988.