

# Discovering a Language for Human Activity

Gutemberg Guerra-Filho, Cornelia Fermuller, Yiannis Aloimonos

Computer Vision Laboratory  
Center for Automation Research  
University of Maryland  
College Park, MD 20742-3275  
[guerra@cs.umd.edu](mailto:guerra@cs.umd.edu), [{yiannis,fer}@cfar.umd.edu](mailto:{yiannis,fer}@cfar.umd.edu)

## Abstract

We present a roadmap to a language for symbolic manipulation of visual and motor information in a sensory-motor system model. The visual information is processed in a perception subsystem which translates a visual representation of action into our visuo-motor language. One instance of this perception process could be achieved by a Motion Capture system. We captured almost 90 different human actions in order to have empirical data that could validate and support our embodied language for movement and activity. The embodiment of the language serves as interface between visual perception and the motor subsystem responsible for action execution. The visuo-motor language is defined using a linguistic approach. In phonology, we define basic atomic segments that are used to compose human activity. Phonological rules are modeled as a finite automaton. In morphology, we study how visuo-motor phonemes are combined to form strings representing human activity and to generate a higher-level morphological grammar. This compact grammar suggests the existence of lexical units working as visuo-motor subprograms. In syntax, we present a model for visuo-motor sentence construction where the subject corresponds to the active joints (noun) modified by a posture (adjective). A verbal phrase involves the representation of the human activity (verb) and timing coordination among different joints (adverb).

## 1. Introduction

Human activity involves breathing patterns, eye movements, postures, limb actions, head movement, trunk movement, change of stance, facial expressions, and orientations in space.

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.

An artificial intelligence with commensurate abilities requires a symbolic structure for reasoning about human activities. A symbolic structure is used in recognition, learning, imitation, and motor planning with a compact representation.

We propose a sensory-motor system model to support activity understanding. In this paper, we concentrate in the

discovery of a visuo-motor language as the symbolic structure for our sensory-motor system model. The visuo-motor language for human activity representation is specified in a linguistic approach, where phonetics, morphology, and syntax are defined. A language for human activity has impacts in many fields.

In kinesiology and movement analysis, the symbolic representation materializes the concept of motor programs and enables the identification of common motor subprograms used in different activities as demonstrated in this paper. This way, the generation of such language allows exploring how a motor activity vocabulary is organized in terms of its subprograms.

Spoken language and visible movement uses a similar cognitive substrate based on the embodiment of grammatical processing. With evidence that language is grounded on the motor system, a visuo-motor language will provide linguistics with a step towards the hypothesis of a universal grammar and allow more development on this subject.

In Computer Vision, a visuo-motor language allows the visual parsing of human movement which may be used in action recognition and video annotation to extract symbolic descriptions from real-world data.

The visuo-motor language may also help humanoid robots to generalize the planning and control of motor activities while using a vocabulary of human actions. In Computer Graphics, this language could be a basic foundation for a different approach in animation programming.

Finally, we propose a visuo-motor language for human activity as an analytical tool. This tool would allow the use of parsing and symbolic reasoning about this kind of information. This representation is compact, computationally more efficient, and consistent with many biological evidences such as mirror neurons and Brocas's region functionality.

Motivation for a linguistic approach comes from the fact that movement patterns are similar to language. In some ways, they are composed of elements in combination and sequences, but they may not be organized like language because the dimensions are qualitatively different in important ways (Armstrong, 1995).

In section 2, we discuss previous work related to symbolic representations of human activity. Section 3 presents our

sensory-motor system model. The embodiment of a visuo-motor language is discussed in Section 4 using a gradual transformation of visual representations and the musculo-skeletal system perspective. The visuo-motor language discovered in this paper is presented in Section 5. Conclusion and future work are addressed in Section 6.

## 2. Previous Work

Symbolic representations of human activity are found in movement notation systems developed for dance and in linguistic studies about gestural 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). 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. 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.

Evidence towards language embodiment grounded in spatio-motor 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).

## 3. Sensory-Motor System

Broca's region in the human brain is related to various functions ranging from *perception* to *action* (Nishitani *et al.*, 2005). An action-perception link in Broca's area involves *learning* (e.g. language and skill acquisition) through *imitation*.

Mirror neurons would activate when a monkey performs a specific action with its hand. The same neurons will also fire when the monkey observes the same action (Gallese *et al.*, 1996). There is evidence that mirror neurons exist in human brains, active during observation and execution of an action. However, Broca's region was not activated when human subjects watched an action that is not in the observer's motor vocabulary (Buccino *et al.*, 2004). This evidence suggests that action *recognition* is another

function related to Broca's area. Broca's region also contributes to action *planning* which is related to prediction.

Action understanding involves mapping observed motor sequences onto a vocabulary of actions. This vocabulary represents motor patterns performed previously and stored according to some knowledge representation.

The Broca's region functionality and the mirror neurons theory suggests that perception and action share the same symbolic structure as a knowledge that provides common ground for recognition and motor planning. We propose a sensory-motor system model with six subsystems (perception, action, learning, imitation, recognition, and motor planning) where our visuo-motor language plays a central role (see Fig. 1).

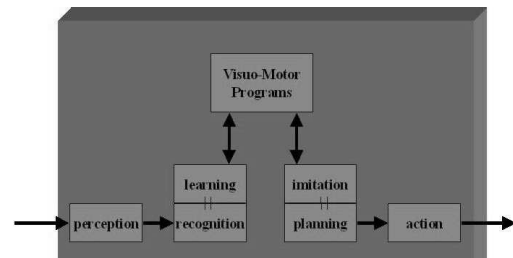


Fig. 1: Sensory-motor system model.

The perception subsystem takes the visual input and extracts higher-level representations for human actions. These representations are parsed and possibly matched to visuo-motor programs by the recognition process. If the action vocabulary does not contain the observed action, no matching is found and learning occurs through imitation. The imitation process takes advantage of the same framework used for action planning and, basically, searches for a physically feasible plan to execute the observed unknown action in the action subsystem.

## 4. Embodiment and Representations

Visual and motor representations for human activity are coupled by embodiment. These two aspects of human activity are abstracted to a common ground which is, ultimately, the consideration of the human body into the modeling process. In this paper, we focus in the discovery of a common embodied symbolic language while the visual and motor abstraction process is only suggested.

### 4.1. 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 the other for movement (Clark, 1963: 274). The visual system is an organization that maps into a simpler 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.

The lowest level of representation (global) captures the whole body motion, while a higher-level representation (structured) 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 (joints) of the actor while simplifies the classification process involved in movement recognition. Empirically, the visual representations range from motion fields to 2D joint angles derived from a stick model of the human body (see Fig. 2).

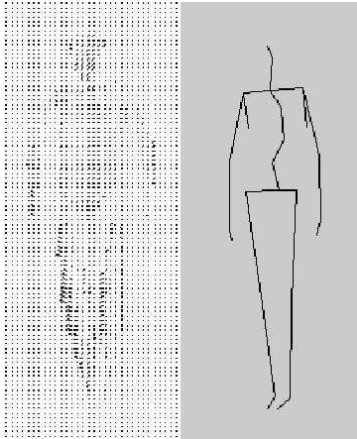


Fig 2: Visual representations from motion field to stick model.

One instance of the visual abstraction process is achieved by a Motion Capture (MoCap) system, where multiple images of a human activity are captured simultaneously. The action is performed in a special environment while the actor wears a specific suit with markers over the body joints. We captured videos featuring 90 different human activities and the corresponding three-dimensional reconstruction for trajectories of body parts was found using our own MoCap system. Given this three-dimensional reconstruction, joint angles were computed to describe human movement (see Fig. 3).

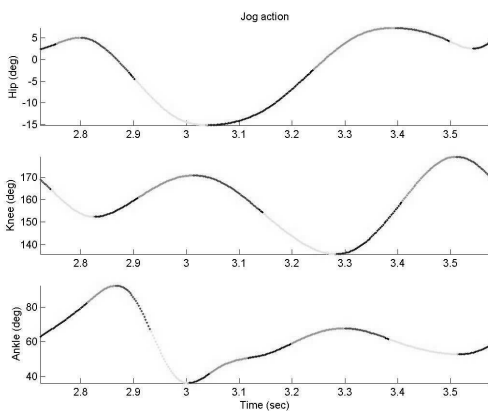


Fig. 3: Joint angles for ankle, knee, and hip during jog activity.

Relaxing the assumptions used in a MoCap system (multiple cameras, special environment, specific suit)

towards a monocular system in a more general environment, the abstraction process has to transform global 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. These constraints are applied to each representation in order to get more structure into a higher-level representation. One example of a constraint considers the movement of points in the same body part as rigid. A more abstract constraint may use the topological connections of body parts.

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

## 4.2. 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 seconds (single twitch). In order to perform longer smooth controlled muscular contractions, the muscle needs to be stimulated repeatedly. Your brain will send a stream of impulses through a certain number of nerves to the muscle in order 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.

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 cognitive process. An anatomic action corresponds to a subset of the motor state. For example, the elbow→extends action corresponds only to the activation

of the muscles anconeus, brachioradialis, and triceps brachii.

Anatomic actions can be divided into flexion (muscle contraction causing a bending movement), extension (straightening or a return from flexion), and rotation. Most movement patterns are a combination of these three muscle movements. In general, an anatomic action performed by a specific joint and occurring in a particular anatomic plane (transverse, frontal, sagittal) corresponds to a degree of freedom (DOF) in a human body.

A joint angle time-varying function for all DOF represents a human action. This approach to action definition is used in this paper for the derivation of a symbolic representation.

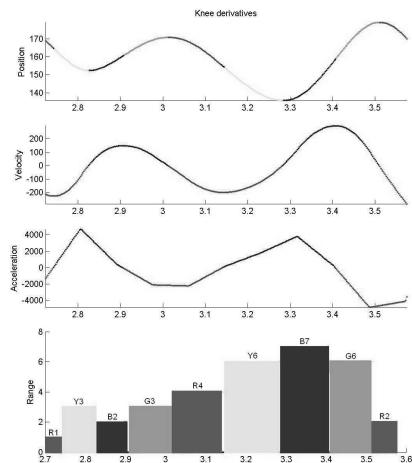
## 5. Visuo-Motor Language

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). Words represent classes of things and actions. Sentences represent very basic relationships between words.

### 5.1. Phonology

Phonology is the system that selects certain states among all possible states in the system and presents them as phonemes, the segments composing words. In a visuo-motor system, a phonetic description consists of a sequence of static physical measures (articulatory configurations).

A phonetic representation is a characterization of how a physical system changes over time (Browman and Goldstein, 1985: 35). This way, our visuo-motor phonology is based on first derivatives (angular velocity) and second derivatives (angular acceleration) of joint angles (see Fig.



4).

Fig. 4: Knee angle derivatives during jog activity.

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. This supports an evolution of language ability from the motor system (Kien, 1992: 19).

The joint movement is segmented according to the sign of the angular derivatives. This way, the six possible atomic states are the four combinations of positive and negative signs for the two derivatives, one state for zero acceleration, and another state for zero velocity. The first segment in the knee joint during the jog action has negative velocity and negative acceleration.

Each segment has an initial joint angle  $\varphi$  and a final angle  $\theta$ . The potential  $\Delta$  of a segment is the discrete absolute difference between these two angles:  $\Delta = \lfloor |\varphi - \theta| / \mu \rfloor$ , where  $\mu$  is a constant. In the first segment of the knee joint movement during the jog action,  $\varphi$  is  $165.43^\circ$  and  $\theta$  is  $168.95^\circ$ , consequently,  $\Delta = \lfloor |165.43^\circ - 168.95^\circ| / 3^\circ \rfloor = 1$ , where  $\mu = 3^\circ$ .

An alphabet of atomic joint movements is necessary for a symbolic representation of action. Each segment corresponds to an atom  $\alpha\Delta$ , where  $\alpha$  is a symbol associated with the segment's state and  $\Delta$  corresponds to the segment's potential. The symbol R is assigned to negative velocity and negative acceleration; the symbol Y is assigned to negative velocity and positive acceleration; the symbol B is assigned to positive velocity and positive acceleration; the symbol G is assigned to positive velocity and negative acceleration, the symbol V is assigned to zero acceleration, and the symbol  $\square$  is assigned to zero velocity. This way, the string R1 Y3 B2 G3 R4 Y6 B7 G6 R2 corresponds to the activity jog according to the knee joint (see Fig. 4).

Only certain combinations of different movement segments constitute the words of human activity. In order to describe accurately the phonology of human movement, phonological rules are required. The rules for our visuo-motor phonology are specified using a finite automaton (see Fig. 5). The transitions in the automaton are based on the observation that changes in the sign of the first and second derivatives do not occur simultaneously.

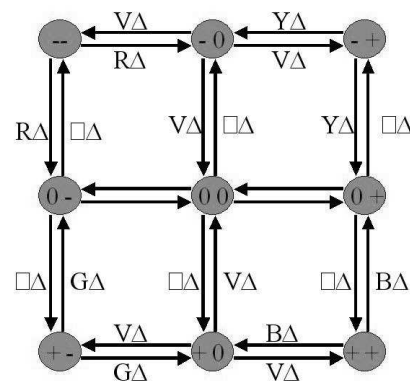


Fig. 5: Visuo-Motor phonological rules.

At the message or mental level, abstract units of language are discrete, static, and context-free. 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.

The realization consequences appear in the discretization of the segment's potential and in the phonological rules. In order to overcome these effects, we consider two atoms  $\alpha\Delta_1$  and  $\alpha\Delta_2$  to be the same if  $|\Delta_1 - \Delta_2| \leq 1$ . We also relax the phonological rules to avoid atoms executed in a very short period of time. This may be implemented using a time length threshold.

## 5.2. Morphology

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.

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 (Perrett, 1989: 109). A complex movement combined with others forms a larger structure (coordinated patterns of gestures in time and space) that defines a word (Kelso, Saltzman, and Tuller, 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 visuo-motor word.

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 (Brownman and Goldstein, 1990: 300).

Given the symbolic representation for an activity lexicon, a hierarchical organization and morphological grammars are derived for the action lexicon. Our experimental lexicon consists in 10 locomotion actions: jog, jump, run, scuff, stomp, swagger, tiptoe, toe, troop, and walk.

```
jog := B0 G0 R2 Y3 B3 V0 G3 R0 Y0
jump := G0 R0 Y0 B2 G1 V0 B5 G3 R0 Y0 R10 Y5 B0 G0 R0 Y0 B0 G1 B0 G0 R0 Y1 B0 V0 B0 V0 G0 R0 Y0 B0 G0 R0 Y0 V0
run := B4 G7 R1 Y2 B1 G0 R3 Y5 V0 Y0
scuff := R0 V0 Y0 V1 Y0 V0 R0 V0 Y0 B0 V0 G0 V1 G0 V0 G0 : 0 V0 : 0
stomp := G0 R0 Y0 V0 R1 V0 Y2 V1 R0 V0 Y1 V0 Y0 B0 G0 V0 B1 V0 G1 V0 B1 V0 G2 R2 Y0 B0
swagger := Y0 V0 R1 V0 Y0 V0 R0 V0 Y0 V0 : 0 R0 V0 Y0 V0 R0 V0 Y0 B2 G2 V0 B0 V0 G0 V0 B0 V0 G0 V0 G0 R1
tiptoe := B0 G1 V1 B0 V0 G1 : 0 B0 V0 G0 R0 V0 Y0 Y0 Y0 V0 R0 V0 Y0 B0 V0 G0 R0 V0 Y0
toe := R0 V1 R0 V0 Y0 R0 Y0 B0 G1 V2 G1 R0 V0 Y0 B0 G0
troop := R0 Y0 B0 G0 R1 V3 Y3 V1 Y0 B5 G3 R0 Y0 B2 G3 R2 V0 Y2 B0 G0
walk := R0 V0 Y0 V1 Y0 V0 R0 V0 Y0 B2 G2 V0 B0 V0 G0 R0 V0 Y0 : 0 R0 Y1 V0 R0 V0 Y1
```

Fig. 6: Morphological grammar at the lowest level.

In this experiment, the morphological grammar generation considers only the right hip joint without loss of generality. The visuo-motor strings for each action word represent the lowest level in the morphological grammar (see Fig. 6).

The generation of a higher-level morphological grammar involves finding common substrings in different actions of the lexicon. Our algorithm finds the most frequent pair  $\alpha_i\Delta_i$ ,  $\alpha_j\Delta_j$  of consecutive atoms in the current grammar. In our example, for the morphological grammar at the lowest level, the most frequent consecutive pair of atoms is V0 Y1 with 25 occurrences. A new grammar rule  $L_n := \alpha_i\Delta_i \alpha_j\Delta_j$  is then created. This way, the first rule created in our experiment is L01 := V0 Y1. A higher-level grammar is generated using the new rule. Each occurrence of the pair of atoms  $\alpha_i\Delta_i$ ,  $\alpha_j\Delta_j$  (V0 Y1) in the current grammar is replaced by a non-terminal  $L_n$  (L01). The second level grammar for our example has one more non-terminal but the rules are more compact (see Fig. 7). This process is repeated until the most frequent pair in the current grammar has less than two occurrences and, consequently, the highest level of the grammar is reached.

```
L01 := V0 Y1
jog := B0 G0 R2 Y3 B3 V0 G3 R0 Y0
jump := G0 R0 Y0 B2 G1 V0 B5 G3 R0 Y0 R9 Y5 B0 G0 R0 Y0 B0 G1 B0 G0 R0 Y1 B0 V0 B0 V0 G0 R0 Y0 B0 G0 R0 Y0 V0
run := B4 G7 R1 Y2 B1 G0 R3 Y5 L01
scuff := R0 L01 L01 V0 R0 L01 V0 R0 L01 B0 V0 G0 V1 G0 V0 G0 : 0 V0 : 0
stomp := G0 R0 Y0 V0 R1 L01 V1 R0 L01 L01 B0 G0 V0 B1 V0 G1 V0 B1 V0 G2 R2 Y0 B0
swagger := Y0 V0 R1 L01 V0 R0 L01 V0 : 0 R0 L01 V0 R0 L01 B2 G2 V0 B0 V0 G0 V0 B0 V0 G0 V0 G0 R1
tiptoe := B0 G1 V1 B0 V0 G1 : 0 B0 V0 G0 R0 L01 L01 V0 R0 L01 B0 V0 G0 R0 L01
toe := R0 V1 R0 L01 V0 R0 Y0 B0 G1 V2 G1 R0 L01 B0 G0
troop := R0 Y0 B0 G0 R1 V3 Y3 L01 B5 G3 R0 Y0 B2 G3 R2 L01 B0 G0
walk := R0 L01 L01 V0 R0 L01 B2 G2 V0 B0 V0 G0 R0 L01 : 0 R0 Y1 V0 R0 L01
```

Fig. 7: Morphological grammar at the second level.

The highest level of the grammar contains the lexical units (words) of a visual-motor language. The sub-string alphabets embed the structure that allows the identification of roots, prefixes, and suffixes in the lexical units. Furthermore, this structure implies relations that give rise to a hierarchical organization (see Fig. 8).

```
L01 := V0 Y1
L02 := R0 L01
L03 := B1 G1
L04 := V1 G1
L05 := R1 Y1
L06 := V0 L02
L07 := B0 L04
L08 := L05 L03
L09 := V0 L07
L10 := L03 L09
L11 := L01 L06
L12 := L05 L06
L13 := L02 L11
L14 := L08 R2
L15 := V0 -0
L16 := B5 G3
```

L17 := L16 L05  
 L18 := L09 L08  
 L19 := L12 L06  
 L20 := L14 Y4  
 jog := B3 V0 G3 L20  
 jump := L08 V0 L17 R9 Y5 L03 L08 L03 L05 B0 L18 L05 L04  
 run := B4 G7 L20 L01  
 scuff := L13 L06 L07 L04 L04 □0 L15  
 stomp := L19 L01 L10 L18  
 swagger := L15 L02 L06 L10 L09 L04 L19  
 tiptoe := L10 □0 L07 L13 L07 L02  
 toe := R0 L06 V0 L08 L04 L02 L03  
 troop := L14 V3 Y3 L01 L17 B2 G3 R2 L01 L03  
 walk := L13 L10 L02 □0 L12

Fig. 8: Morphological grammar at the highest level.

The generated grammar is an effective way to find common subprograms in the activity lexicon. In the highest level grammar of our experiment, the non-terminal L13 belongs to the activities scuff, tiptoe, and walk. This way, the string L13  $\equiv$  R0 V0 Y1 V0 Y1 V0 R0 V0 Y1 is a motor subprogram used in those actions (see Fig. 9).

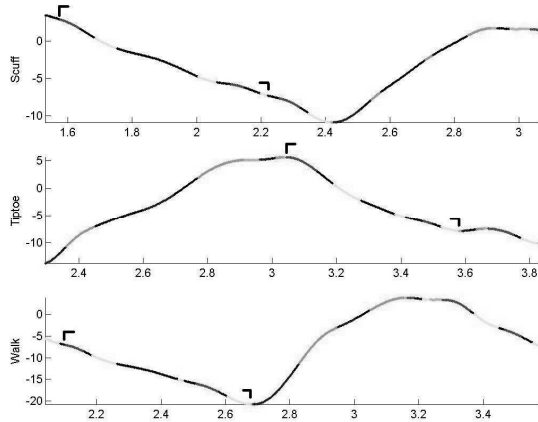


Fig. 9: A common motor subprogram.

### 5.3. Syntax

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.

A word and a sentence often look identical in sign language. Different kinds of segment morphemes combine to form word sentences: posture and activity. A posture describes how articulatory features (moveable parts) are configured. Activity is defined as a period of time during which some aspect of articulation is in transition.

The Subject-Verb-Object (SVO) pattern of syntax is a reflection of the patterns of cause and effect: something

doing something to something else. Linguistic expressions are processed as if they were objects with internal structural configurations (Deane, 1991). 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 words, the action is transitive and involves an object or another patient body part.

The subject in a visuo-motor sentence corresponds to a binary string specifying which body parts are active. Usually, a locomotion activity involves lower limbs joints. Therefore, in our experiments, the subject is represented by a binary string signaling these joints.

The initial posture of the sentence is analogous to an adjective which further describes the subject. For each joint  $j$  in the subject, the visuo-motor adjective corresponds to a non-negative integer  $p_j$  computed as  $\lfloor (\phi - \theta_{min}) / \sigma \rfloor$ , where  $\phi$  is the initial joint angle,  $\theta_{min}$  is the minimum value in the joint angle range, and  $\sigma$  is a constant. The initial posture for our experimental lexicon is represented by strings considering only the active joints in the lower limbs (see Fig. 10).

	jog	jump	run	scuff	stomp	swagger	tiptoe	toe	troop	walk
R_Hip	4	4	2	4	5	2	1	3	4	2
R_Knee	7	7	4	8	6	8	8	9	7	9
R_Ankle	8	5	4	5	4	3	0	3	4	4
L_Hip	3	3	4	2	3	1	3	2	2	2
L_Knee	1	9	7	9	7	8	9	7	6	7
L_Ankle	5	4	7	4	3	0	4	0	5	3

Fig. 10: Visuo-motor adjectives for our experimental lexicon.

The sentence verb represents the changes each active joint experiences during the action execution. The representation for a visuo-motor verb was discussed in the previous subsections. However, further description is required to deal with coordination among different joints.

For synchronization of different joints performing in the same action, each atomic segment is associated with the corresponding discrete version of the time interval length. This way, the visuo-motor strings representing each joint movement (verb) are augmented by the timing. The string for the knee joint in the jog action becomes R1<sub>7</sub> Y3<sub>30</sub> B2<sub>28</sub> G3<sub>39</sub> R4<sub>47</sub> Y6<sub>51</sub> B7<sub>45</sub> G6<sub>38</sub> R2<sub>25</sub>, where the subscript specifies time length of execution.

A coordinated segment is a time interval delimited by events representing local minima and maxima in the joint angle function for any of the active joints. These events occur in between specific atomic pairs ( $Y\Delta$   $B\Delta$  and  $G\Delta$   $R\Delta$ ) and, consequently, may be computed from the augmented visuo-motor verb strings. For the jog action, there are 20 coordinated segments (see Fig. 11).

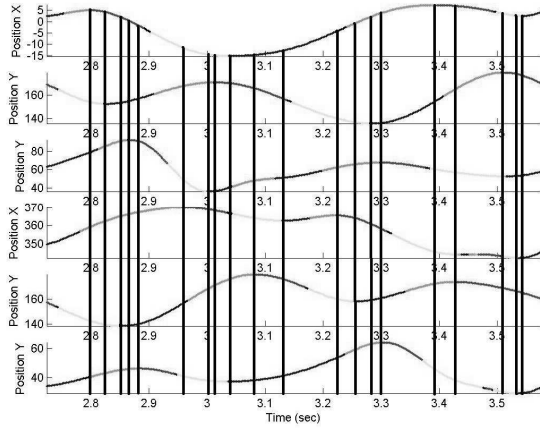


Fig. 11: Coordinative segments for the jog action.

A visuo-motor adverb is a string of values (multiplicative constants) in the range  $[1-e, 1+e]$ , where  $e$  is an elasticity constant that specifies a bound for the variation in the execution time of each coordinated segment. The adverb string is discretized such that each value  $v$  corresponds to an integer  $\lfloor (v - (1 - e)) / \rho \rfloor$ , where  $\rho$  is a constant. A visuo-motor adverb is appended to a verb in such a way that each value in the adverb string corresponds to a coordinated segment in the verb. A visuo-motor sentence  $S$  consists of noun phrase (noun + adjective) and verbal phrase (verb + adverb):  $S := NP VP$ , where  $NP := N Adj$  and  $VP := V Adv$  (see Fig. 12).

	Subject (Noun)	Modifier (Adjective)	Predicate (Verb)	Modifier (Adverb)
	lower limbs	standing	jog	slowly
R_Hip	1	4	B0 <sub>10</sub> G0 <sub>16</sub> R2 <sub>34</sub> Y3 <sub>30</sub> B3 <sub>67</sub> V0 <sub>8</sub> G3 <sub>35</sub> R0 <sub>37</sub> Y0 <sub>16</sub> B0 <sub>13</sub>	
R_Knee	1	7	R1 <sub>1</sub> Y3 <sub>30</sub> B2 <sub>28</sub> G3 <sub>39</sub> R4 <sub>47</sub> Y6 <sub>61</sub> B7 <sub>63</sub> G6 <sub>58</sub> R2 <sub>25</sub>	
R_Ankle	1	8	B6 <sub>33</sub> G3 <sub>39</sub> R8 <sub>83</sub> Y9 <sub>26</sub> B1 <sub>14</sub> G2 <sub>27</sub> B2 <sub>28</sub> G2 <sub>27</sub> R1 <sub>29</sub> Y2 <sub>48</sub> B1 <sub>23</sub>	
L_Hip	1	3	B2 <sub>20</sub> G4 <sub>66</sub> R0 <sub>33</sub> Y1 <sub>32</sub> B0 <sub>16</sub> G0 <sub>16</sub> R2 <sub>33</sub> Y3 <sub>33</sub> R0 <sub>34</sub> Y0 <sub>11</sub> B0 <sub>13</sub>	34343324333423434423
L_Knee	1	1	Y4 <sub>40</sub> B6 <sub>62</sub> G6 <sub>63</sub> R3 <sub>33</sub> Y3 <sub>30</sub> B2 <sub>31</sub> G2 <sub>31</sub> R4 <sub>56</sub>	
L_Ankle	1	5	B1 <sub>21</sub> G1 <sub>31</sub> R1 <sub>23</sub> Y1 <sub>32</sub> B6 <sub>63</sub> G2 <sub>23</sub> R4 <sub>26</sub> Y6 <sub>39</sub> Y0 <sub>13</sub> B1 <sub>16</sub>	

Fig. 12: A visuo-motor sentence for the jog activity.

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 in order to produce an action and to combine actions into larger structures: words. The sequential combination of action sentences must obey the cause and effect rule. The visuo-motor 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  $NP_i + VP_i = NP_{i+1}$  (see Fig. 13). In a higher-level, an example of the cause and effect rule as a syntax rule for action sequencing is **sitting + stand up = standing**. The cause and effect rule is physically consistent and embeds the ordering concept of syntax.

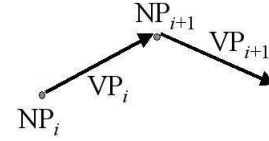


Fig. 13: The cause and effect rule in a vector space.

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. In written language, sentences are delimited by punctuation. Analogously, the action language delimits sentences using motionless actions such as stop, still, and freeze, for example. In order to analyze action syntax, a corpus of sentences is required. 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 lexical unit.

## 6. Conclusions

In this paper, we provide the tools for the construction of a visuo-motor language which is the basic kernel for symbolic manipulation of visual and motor information in a sensory-motor system model proposed. The visual information is processed in a perception subsystem which translates a global visual representation of action into an embodied representation matched to our visuo-motor language. Our instance of the visual process is a Motion Capture system which reconstructs human movement from multi-view images. The embodiment is an important characteristic of the language serving as interface between visual perception and the motor planning subsystem towards action execution. The visuo-motor language is defined using a linguistic approach by specifying phonology, morphology, and syntax of the visuo-motor information.

In phonology, we presented a suggestion for the basic atomic segments that are used to compose human activity. Segments are characterized according to the sign of the first and second angular derivatives of joints. Phonological rules were derived from this characterization and modeled as a finite automaton.

In morphology, we studied how visuo-motor phonemes were combined to form strings representing human activity. Basically, we explored common substrings to generate a higher-level morphological grammar which is more compact and suggests the existence of lexical units working as visuo-motor subprograms.

In syntax, we presented a model for visuo-motor sentence construction where the subject in a sentence corresponds to the active joints (noun) modified by a posture (adjective). This way, in a higher-level of abstraction, an example of a noun phrase would be *lower limbs standing*. A verbal phrase involves the representation of the human activity (verb) according to the phonology and morphology discussed previously. Coordination among different joints is specified by timing the atomic segments and appending elastic discrete values (adverb) to coordinated segments. The adverb is used to adjust and modify the action execution. The highest abstraction of an adverb would be analogous to a verbal phrase such as *jog slowly*, where *jog* is a verb modified by the adverb *slowly*.

For future work, we plan to investigate other morphological grammar generation algorithms and evaluate these algorithms according to the compactness of the language. In this paper, the grammar generation considers each joint independently. We intend to evaluate the grammar generation process when there is no distinction among joints and, consequently, the common subprograms are shared even by different joints in different activities.

The phonology and morphology of nouns, adjectives, and adverbs are issues that deserve more attention. We will search for higher-level standard sets of active joints, postures and coordination control. This way, we may find an abstract way to describe human activity using nouns as *lower limbs and right arm*; adjectives as *standing, kneeling, sitting, and lying*; and adverbs as *fast and slowly*.

## References

Armstrong, D., Stokoe, W., and Wilcox, S. 1995. *Gesture and the nature of language*. New York: Cambridge University Press.

Boyd, J. and Little, J. 1997. Global versus structured interpretation of motion: moving light displays. In Proceedings of IEEE Nonrigid and Articulated Motion Workshop at CVPR 97, 15-16.

Browman, C. and Goldstein, L. 1985. Dynamic modeling of phonetic structure. In V. Fromkin (ed.), *Phonetic linguistics*. New York: Academic Press.

Browman, C. and Goldstein, L. 1990. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics* 18: 299-320.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, CA., and Rizzolatti, G. 2004. Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. *Journal of Cognitive Neuroscience* 16: 114-126.

Clark, W. 1963. *The antecedents of man*. New York: Harper and Row.

Deane, P. 1991. Syntax and the brain: neurological evidence for the spatialization of form hypothesis. *Cognitive Linguistics* 2(4): 361-367.

Dell, C. 1971. *A primer for movement description*. Dance Notation Bureau, Inc., New York.

Edmondson, W. 1987. Segments in signed language: do they exist and does it matter? In 4<sup>th</sup> International Symposium on Sign Language Research, Helsinki.

Eshkol, N. 1980. *50 Lessons By Dr. Moshe Feldenkrais*. Tel-Aviv: The Movement Notation Society.

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

Hutchinson, A. 1977. *Labanotation*. Theatre Arts Books.

Kelso, J., Saltzman, E., and Tuller, B. 1986. The dynamical perspective on speech production: data and theory. *Journal of Phonetics* 14: 29-59.

Kien, J. 1992. Temporal segmentation in the motor system, symbolization, and the evolution of language. Annual meeting of the Language Origin Society, Cambridge, UK.

Lakoff, G. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press.

Lidell, S. 1984. THINK and BELIEVE: sequentiality in American Sign Language. *Language* 60: 372-399.

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

Perlmutter, D. 1988. A mosaic theory of American Sign Language syllable structure. In 2<sup>nd</sup> Conference on Theoretical Issues in Sign Language Research. Gallaudet University, Washington DC.

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* 146: 87-113.

Sandler, W. 1986. The spreading hand autosegment of American Sign Language. *Sign Language Studies* 50: 1-28.

Studdert-Kennedy, M. 1987. The phoneme as a perceptuomotor structure. In D. Allport (ed.), *Language perception and production: relationships between listening, speaking, reading and writing*. London: Academic Press.

Wilbur, R. 1987. *American Sign Language: linguistic and applied dimensions*. Boston: College Hill Press.