Title

Learning a Motor Grammar of Iconic Gestures

Permalink

https://escholarship.org/uc/item/9kh3j9zr

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 36(36)

ISSN

1069-7977

Authors

Sadghipour, Amir
Kopp, Stefan

Publication Date

2014

Peer reviewed

Learning a Motor Grammar of Iconic Gestures

Amir Sadeghipour (sadeghipour@uni-bielefeld.de)

Stefan Kopp (skopp@techfak.uni-bielefeld.de)

Faculty of Technology, Center of Excellence ‘Cognitive Interaction Technology’ (CITEC),
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany

Abstract

In this paper, we present a computational investigation into the compositionality of iconic gestures by trying to learn a motor grammar. We propose a grammar formalism that learns (1) the salient, invariant features of single movement segments (motor primitives) and (2) the hierarchical organization of these segments in complex gesturing. The formalism is applied to a dataset of natural iconic gestures. The extracted structure reveals compositional patterns of iconic gesturing.

Keywords: gesture; motor program; probabilistic grammar

Introduction

An integral part of our communicative ability is to gesture, i.e. to perform expressive bodily actions as (part of) an utterance. Gesturing has received much interest during the last decades and work in (Psycho-)Linguistics, Cognitive Psychology and Human-Computer Interaction has provided many models of gesture production (e.g. Kopp, Bergmann, and Kahl (2013)) or gesture recognition (see Mitra and Acharya (2007) for a survey). Most of these models focus either on the higher cognitive processes (e.g., of multimodal conceptualization or speech-gesture formulation) or on low-level vision-based perception and recognition. Little is known about the sensory-motor representations that underlie and possibly shape those cognitive processes during perception, interpretation and production of gestural behavior.

We focus on natural iconic gestures, which are spontaneously and extemporaneously performed in communication to refer to objects or events by depicting visual-spatial aspects. As illustrated in Figure 1, spontaneous iconic gesturing exhibits a very large variability even when performed for the same object. Extracting the communicative significance of an iconic gesture thus involves (at least) two steps: (1) an *iconic mapping* of physical movement onto visuospatial imagery (e.g. a circular trajectory onto aspects of “roundness”, “size”, or “orientation”); (2) a *referential mapping* of this imagery onto concrete objects or events (e.g. the specific round window being referred to). Here we are concerned with the first iconic mapping only and we want to understand how movements are used to create gestural imagery.

Classically, iconic gestures are assumed to have no linguistic structure with a composition of primitives. Rather, McNeill (2000) has argued for a “global semiosis” of such gesticulations, holding that the meanings of ‘parts’ of a gesture are determined by the meaning of the whole (and not the other way around as in language). On the other hand, it is widely acknowledged that the motor system is organized hierarchically (Hamilton & Grafton, 2007), such that given intentions are mapped onto motor goals that are then refined into structured motor programs and on to motor primitives (Mussa-Ivaldi & Solla, 2004). In line with this view, Flash and Hochner (2005) argued that our motor repertoire can be spanned by combining motor primitives according to syntactic rules. As the motor system is also involved in the perception of gestures (Montgomery, Isenberg, & Haxby, 2007), this hierarchical structure of motor knowledge should also guide the interpretation of communicative gestures. Against this background, the question is (1) whether and how a compositional and hierarchical motor structure can be identified in gestural movement and (2) how this may guide the comprehension and production of iconic gestures with their global semiosis.

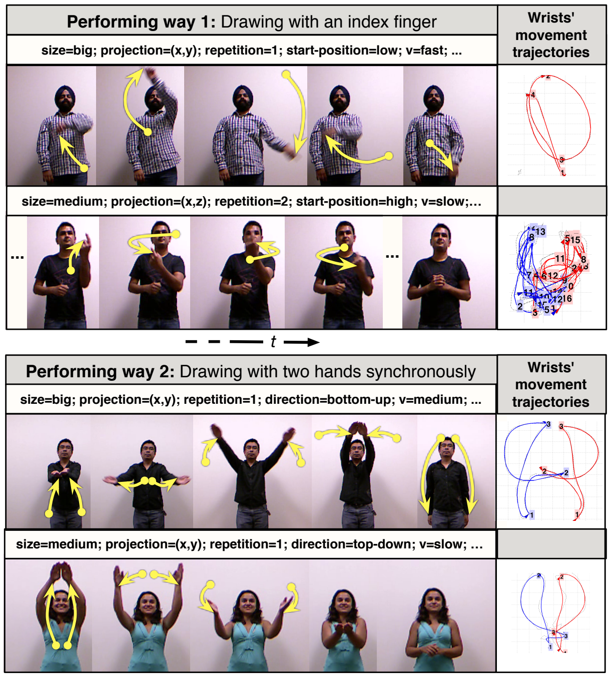

Performing way 1: Drawing with an index finger
Wrists' movement trajectories
size=big; projection=(x,y); repetition=1; start-position=low; v=fast; ...
size=medium; projection=(x,z); repetition=2; start-position=high; v=slow;...
...
t
Performing way 2: Drawing with two hands synchronously
Wrists' movement trajectories
size=big; projection=(x,y); repetition=1; direction=bottom-up; v=medium; ...
size=medium; projection=(x,y); repetition=1; direction=top-down; v=slow; ...

Figure 1: Different iconic gestures performed for a virtual 3D sphere, along with the respective wrist trajectories.

In this paper we present a computational investigation into the compositionality of iconic gesturing by trying to learn a “motor grammar” of iconic gestures. After discussing relevant work, we propose a hybrid grammar-based approach that statistically identifies low-level feature-based regularities (“symbolization process”) and integrates this with searching for their compositional organization in high-level syntactic rules. Then, we discuss results from an application of this formalism to a dataset of natural iconic gesture trajectories. We show how movement features that are characteristic of the iconic mapping are determined by their position within syntactic structure while, the other way around, terminals and syntactic rules emerge from these recurring low-level movement features. It is demonstrated that this two-way interaction of two levels of analysis (identifying primitives and finding compositions) allows for carving out a motor grammar of iconic gesture trajectories.

Background and Related Work

One motivation behind the present work is to develop a cognitive model of the sensorimotor processes involved in the production and perception of meaningful nonverbal behaviors, in particular communicative gesture. Previously, we proposed a hierarchical model of sensorimotor knowledge of hand gestures that consists of three levels of abstraction (Sadeghipour & Kopp, 2011): *motor commands (MC)* controlling segments of a movement trajectory (corresponding to motor primitives), *motor programs (MP)* representing complete gesture performances, and *motor schemas (MS)* capturing the variant and invariant features of the gesture performances that can be employed to fulfill certain communicative functions. Accounting for a dual-use of this motor knowledge for both perception and production of gestures, we have proposed an algorithm for Bayesian belief propagation in-between those levels and have utilized it to enable resonance-based perception, generation and imitation with an embodied virtual agent.

The investigation presented here originally aimed to provide a representation for motor schemas that group individual gestural performances (e.g. of "wiping") together and extract the essence of a certain "gesture class" according to the relevant features for its iconic mapping. This links to recent debates about the analogies between the deeper hierarchical organization of human action and the syntax of language (Pastra & Aloimonos, 2012) as well as possible common underlying neural bases (Mussa-Ivaldi & Bizzi, 2000). Indeed, recursive organization of motor primitives can be found in the representation of different motor knowledge in the human brain (Mussa-Ivaldi & Bizzi, 2000; Flash & Hochner, 2005).

Grammar-based formalisms have been applied for the computer-based recognition of nonverbal behavior (Hong, Turk, & Huang, 2000) or complex activities (Ryoo & Aggarwal, 2006; Kitani, Sato, & Sugimoto, 2006). To deal with uncertainty due to noisy sensors or vision, such approaches have been extended to include probabilities. Stochastic Context-Free Grammars (SCFG) (Stolcke, 1994) where applied for vision-based hand gesture recognition (Ivanov & Bobick, 2000) or activity recognition during a Tower of Hanoi task (Minnen, Essa, & Starner, 2003) based on predefined grammars. Only few approaches have attempted to learn probabilistic grammars for activity recognition in domains like gymnastic exercises, traffic events or multi-agent interactions (Kitani et al., 2006; Zhang, Tan, & Huang, 2011). All of them presume a set of clear-cut morphemes and then built up a grammatical structure on top of them (see, e.g., Guerra-Filho and Aloimonos (2007) for the use of "action morphemes"). The challenge here is that we cannot make such an assumption for gesture. Instead we need to ground the learning of hierarchical motor structure in the lowest levels of single feature values, such that structural-syntactic patterns of gestures and low-level statistical regularities that may constitute building blocks are extracted at the same time. To this end, we adopt an idea that can be traced back to the 80's, where Tsai and Fu (Tsai & Fu, 1980) proposed grammars that integrate statistical consideration into syntactic pattern analysis. However, many applied grammar formalisms in behavior analysis have treated these separately with a first step of statistical segmentation and symbolization, usually using Hidden Markov Models (Chen, Georganas, & Petriu, 2008; Ivanov & Bobick, 2000) to learn a finite set of symbols as terminals, and afterwards extraction of longer range structures using grammar formalisms such as SCFG.

Feature-Based Stochastic Context-Free Grammar

We propose *Feature-based Stochastic Context-Free Grammars (FSCFG)* to unify statistical (feature-based) and syntactic (structure-based) processing in a hybrid framework. It aims to learn not only rule-based syntactic compositions of terminals but also the statistical relations in underlying features spaces that form them. Here, we explain briefly how an FSCFG is learned from given sample sequences; see Sadeghipour and Kopp (2014) for more details.

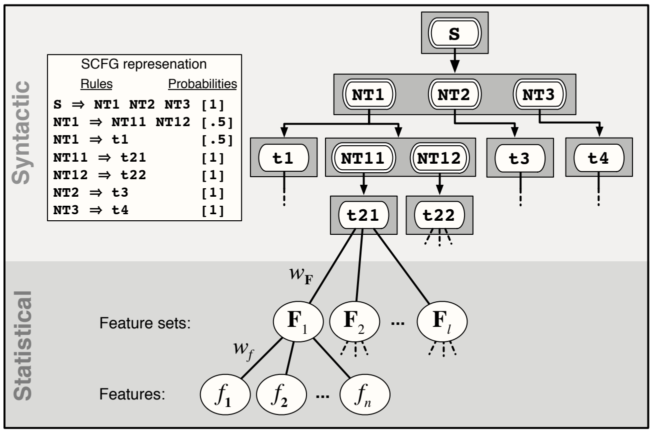


Figure 2: Hybrid model of FSCFG, where terminal symbols in the syntactic structure serve as interface to the statistical feature-based representations.

As illustrated in Figure 2, FSCFG extend the syntactic rules of SCFG to feature-based representation of terminals. In this way, a terminal is not an atomic symbol anymore, but it is represented as a prototype of a cluster of samples in an n-dimensional feature space. Thereby, the impact of the i-th sample (or *feature set*) to a terminal is given by $w_{\mathbf{F}_i}$; and the importance of the i-th feature for each terminal is given by w_{f_i}. Learning these statistical regularities between features and feature sets, the terminals emerge as prototypes of an integrated symbolization process. This allows us to compute the similarity between two symbols (or terminals) as the distance in an n-dimensional feature space. Hence, while parsing, the match between a terminal against an input symbol is not a binary decision but computed probabilistically.

Learning the structure and parameters

Learning an FSCFG refers to optimizing both its structure (i.e. the set of rules), and the values of its parameters (i.e. P of each rule, $w_{\mathbf{F}}$ and w_f for each terminal).

In order to find the optimal set of rules, first an initial set is generated and used to parse the given input strings. For each given input symbol, a terminal with a single feature set is generated and a start rule is added that comprises the entire given string. Upon initialization, the structure is generalized through applying the *merge* and *chunk* operators: The *merge* operator $merge(X_1, X_2) = Y$ replaces all occurrences of the non-terminals X_1 and X_2 with a new non-terminal Y. The *chunk* operator $chunk(X_1 \dots X_k) = Y$ replaces all occurrences of the ordered sequence of the non-terminals $X_1 \dots X_k$ with a single new non-terminal Y. These operators simplify the grammar through decreasing its Description Length (DL) (Rissanen, 1983), which is proportional to the number of bits needed to store the grammar's rules. The *loss measure* during this process is the negative logarithm of the Bayesian posterior parsing probability. The likelihood term, which indicates how well the given samples fit the learned model, is set to the parsing probability of the samples, and the prior probability of a grammar is set to its DL. Using this Bayesian loss measure, the grammar structure is modified towards a trade-off between simplicity and fitting of the model to the given data.

The parameters of an FSCFG are learned and optimized during both learning the structure and parsing new strings. The probability of each rule, P, is determined from how often the rule is invoked in parsing, normalized by the sum of all invocations of rules with the same left-hand side non-terminal. Computing the weight of each feature set of a terminal, $w_{\mathbf{F}}$, employs a counter normalized by the sum of its values in each terminal. Initially, each terminal consists of a single feature set with its counter set to one, yielding $t = \{(\mathbf{F}_1, w_{\mathbf{F}_1} = 1)\}$. This representation of a terminal is adapted in two cases: (1) During parsing, when a terminal parses a symbol, the feature set of the symbol is added to the terminal. In this way, parsing reshapes the terminals of a grammar towards the features of the parsed symbols. (2) During structure learning, when merging two lexical non-terminals $merge(X_1, X_2) = Y$ with $X_1 \Rightarrow t_1$ and $X_2 \Rightarrow t_2$, the right-hand side terminals are merged together yielding a new rule of the form $Y \Rightarrow t$, where $t = \{(\mathbf{F}_1, \frac{1}{2}), (\mathbf{F}_2, \frac{1}{2})\}$. These extensions of feature sets yield an integrated symbolization process in which incremental clustering during both learning and parsing may result in different symbols depending on their syntactical roles in grammar.

The weights of features, w_f, are set for each terminal individually. w_f is defined inversely proportional to the standard deviation of the feature f in all the feature sets of its terminal. In other words, the higher the variance of a feature within a terminal is, the less it contributes to the parsing of that terminal. That means, during parsing, each lexical non-terminal is most sensitive to its most stable features. In this way, FSCFG distinguishes between *variant* and *invariant* features of each terminal, depending on its incorporated feature sets which in turn depend on the structure of the syntactic rules.

Dealing with uncertain input

During grammar learning and parsing, uncertainty in the input data may lead to *deletion errors* when an expected symbol is missing in the input stream, *insertion errors* when symbols occurr spuriously, or *substitution errors* when a symbol is parsed through a wrong terminal. Since FSCFG can parse any symbol by any terminal through feature-based parsing, the substitution error is handled implicitly. To deal with the insertion and deletion errors during parsing, we introduce two new special symbols: *skip* and ε. *skip* is a special terminal with no feature set that handles insertion errors. The probabilistic similarity between *skip* and any given symbol is set to a small value, so that the *skip* terminal will only then be used when otherwise the parsing would fail. On the other hand, ε is an empty terminal which is produced alternatively by each lexical rule with a small probability. Parsing a symbol with ε means to ignore the symbol. From parsing to structure learning, deletion and insertion errors exchange their roles as cause and effect. Correspondingly, using these error handling symbols, the FSCFG framework considers both hypotheses that either the learned grammar structure is incorrect or the given input string is noisy. However, after providing enough training data, noisy rules will be used less for parsing and they will end up with very low probability.

Results

Data analysis

We have collected a dataset of 1739 gestures performed by 29 participants to refer to 20 different 3D objects (3D Iconic Gesture Dataset; 3DIG[1]). Since we aim for investigating the iconic mapping of gestures onto abstract forms such as round or rectangular, we used only the gesture performances which were performed for ten simple geometrical shapes (see Fig. 5, bottom, for objects and Fig. 1 for gesture examples).

A data analysis revealed four major *representation techniques* to perform iconic gestures: (1) *Drawing* the 3D or 2D contour of the object in the air. (2) *Enacting* an action on an imaginary object. (3) *Static posturing* depicts the form of an object with held hands. (4) *Dynamic posturing* refers to a drawing movement with expressive static hand postures. As shown in Figure 3, most of the gestures are performed through drawing or dynamic posturing, where the wrist movement trajectories bear the main contribution to the depiction. To define and represent motor primitives, we hence focused on the wrist trajectories, which also can be easily captured

[1] http://projects.ict.usc.edu/3dig

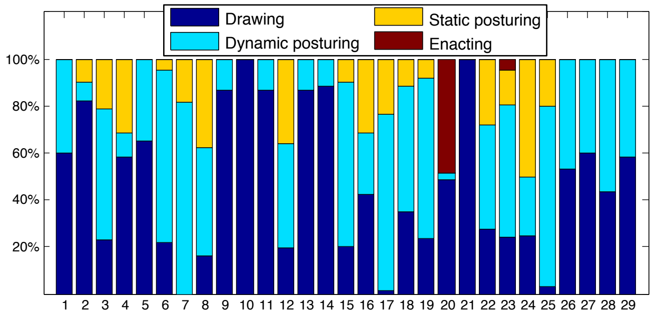


Figure 3: Different techniques used by each participant.

through low-cost motion tracking systems (in contrast to hand shapes).

Table 1: Structural syntactic variabilities of gestures.

Structural variability	**An example of variation**
Degree of simplification	Drawing a 3D shape or a 2D projection.
Ordering	First, referring to the triangle shape of a cone and then to its circular bottom, or vise versa.
Repetition	Drawing a part once, twice or three times.
Handedness	Drawing a circle with one hand or both hands.

We learn the hierarchical motor knowledge as a single FSCFG model from the different gesture performances in the data. Based on our analysis, two types of variabilities need to be accounted for. Table 1 reports the most prominent structural (or syntactic) variabilities, which can lead to very different gestures performed for the same object (i.e. with similar iconic mapping). These variabilities concern segments of a gesture performance. Below this level of variation, the statistical variabilities concern the instantaneous level of spatiotemporal features (Table 2). At this level, features can be *invariant* or *variant* with respect to a specific part of a gesture, to the referred object, or to the used representational technique. For instance, a curved trajectory for round objects is a characteristic and invariant feature, whereas the movement direction of drawing is irrelevant to its iconic mapping and thus a variant feature. Our FSCFG framework needs to cope with both kinds of variabilities while generalizing over different gestures, separating different gesture performances and determining variant and invariant features of each part of a gesture.

Another challenging structural variation of gestures follows from their temporal anatomy. Kendon (1972) divided a gesture performance in three main phases of preparation, stroke and retraction, whereby only the stroke phase is the communicative, expressive part. However, even during this stroke phase, parts of a movement might not be communicative (e.g. when moving the hand to continue drawing from a different position). We observed that the participants moved their hands during non-communicative parts faster than in other parts.

Table 2: Feature-based statistical variabilities of gestures.

Feature-based variability	**An example of variation**
Direction	Drawing while moving a hand to left or to right.
Velocity	Moving a hand fast or slow.
Size	Drawing a small or a big circle.
Position	Gesturing in front of head or chest.
Projection	Drawing a horizontal or vertical projection.
Form	Making a curved movement or a straight one.

Data preparation

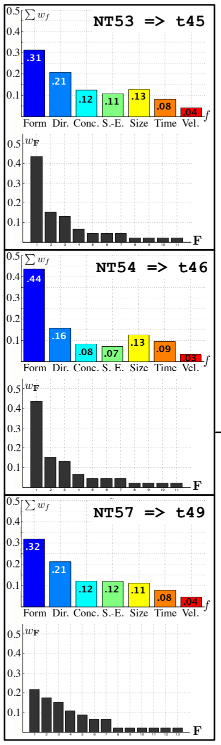
In order to learn FSCFG models of the recorded gestures, the first step is to segment the movement trajectories into a sequence of sub-movements which correspond to individual motor primitives (motor commands in our proposed hierarchy of motor knowledge, encoded here as feature-based symbols of the FSCFG). These motor primitives are demarcated by drops of movement velocity along hand movement trajectories. This segmentation results in short curved or straight movement segments, and a gesture is defined as a string of these segments or symbols. We extracted the following 18 feature dimensions to describe each segment:

- A vector from the start position to its end (3 dimensions)
- Weight and hight of the bounding box (2 dimensions)
- Normal vector of the trajectory's plain (3 dimensions)
- Direction of concavity (3 dimensions)
- Average movement velocity (1 dimension)
- Start timestamp (1 dimension)
- Five samples of the movement trajectory, after normalizing its position, size and orientation (5 dimensions)

Learned motor grammar

In order to learn an FSCFG from observed gestures, the first observation is incorporated through generating new rules. Then, each further gesture performance is first tried to be parsed. If the parsing probability is too low, new rules are added to the grammar to incorporate the performance. After each incorporation, the structure of the FSCFG is optimized (see Learning the structure and parameters) to achieve a generalized sparse representation of the observed gestures.

Figure 4 shows an FSCFG, learned from 211 drawing gestures. The grammar consists of seven start rules, whereby each provides an abstract representation of all gestures with similar iconic mapping. Each start rule is associated with a set of rules, which represent different ways of performing the same iconic mapping with left hand and right hand movement, respectively. Each of these rules consists of nonterminals, each producing a terminal that represents a segment of the wrist movement trajectory. Correspondingly, each terminal is represented statistically, as a set of weighted feature sets (through $w_{\mathbf{F}}$), and each feature set is represented by



```
S => L26 R27 (30)[0.22]
 L26 => MG46 NT5 NT6 NT7 NT8 (17)[0.77]
 L26 => NT17 NT18 NT19 NT20 (1)[0.04]
 L26 => NT18 (1)[0.04]
 L26 => NT5 NT6 NT7 NT8 (3)[0.14]
 R27 => NT10 NT11 NT12 NT13 NT14 NT15(1)[0.04]
 R27 => R27 SK106 (13)[0.59]
 R27 => NT22 NT23 NT24 NT25 (6)[0.27]
 R27 => NT25 (1)[0.04]
 R27 => NT12 (1)[0.04]
S => L31 R37 (17)[0.12]
 L31 => NT27 NT28 NT29 NT30 (16)[0.94]
 L31 => NT28 (1)[0.06]
 R37 => NT32 NT33 NT34 NT35 NT36 (15)[0.75]
 R37 => R37 SK39 (1)[0.05]
 R37 => NT32 NT33 NT35 NT36 (3)[0.15]
 R37 => R37 SK45 (1)[0.05]
S => L54 R54 (3)[0.022]
 L54 => MG46 NT41 (1)[0.33]
 L54 => NT46 NT47 NT48 (2)[0.67]
 R54 => NT43 NT44 (1)[0.33]
 R54 => NT50 NT51 NT52 (2)[0.67]
S => L57 R61 (46)[0.34]
 L57 => NT53 NT54 NT55 NT56 NT57 (2)[0.04]
 L57 => NT53 NT54 NT56 NT57 (6)[0.12]
 L57 => NT53 NT54 NT57 (40)[0.83]
 R61 => NT58 NT59 NT60 (46)[1.0]
S => L70 R76 (14)[0.10]
 L70 => NT64 NT65 NT66 NT67 NT68 NT69(5)[0.31]
 L70 => NT64 NT65 NT66 NT67 NT69 (7)[0.44]
 L70 => NT64 NT65 NT66 NT69 (4)[0.25]
 R76 => NT71 NT72 NT73 NT74 NT75 (10)[0.37]
 R76 => R76 SK77 (12)[0.44]
 R76 => NT71 NT73 NT74 NT75 (5)[0.18]
S => L83 R88 (10)[0.07]
 L83 => NT79 NT80 NT81 NT82 (10)[0.91]
 L83 => L83 SK91 (1)[0.091]
 R88 => NT84 NT85 NT86 NT87 (8)[0.88]
 R88 => NT87 (1)[0.11]
S => L97 R102 (17)[0.12]
 L97 => NT93 NT94 NT95 NT96 (14)[0.93]
 L97 => NT94 (1)[0.07]
 R102 => NT98 NT99 NT100 NT101 (15)[0.94]
 R102 => NT99 (1)[0.06]
```

Figure 4: Motor grammar learned from 211 gesture performances (lexical rules omitted) along with the weights of the features and feature sets of three terminals used in one specific rule.

18 features which are weighted through w_f.

Figure 4 shows the weighting of the terminals in a specific rule representing the movement of the left hand while drawing a semicircle. The terminal produced by NT53 and NT57 represent preparation and retraction phase of gestures respectively, and NT54 represents the actual stroke phase with a drawing intention behind it. A significant difference between this phase and the neighbouring ones is the importance of the form features, indicated by the relative difference between their weigths (blue bar) and those of the other features. Apparently, form features are the most invariant for the stroke phase of drawing gestures for round shapes. On the other hand, in NT53 and NT57, features such as direction, start-end positions and concavity gain more importance. This is due to their low variance as preparation and retraction phases are fast movements in specific directions.

While w_f determines how characteristic each feature for the feature sets of each terminal is, $w_{\mathbf{F}}$ (given as gray bars in Fig. 4) specifies the importance of each feature set itself to its terminal. Few highly weighted feature sets (such as for NT53 and NT54) indicate a prototypical movement segment for that part of the gesture. In contrast, several equally weighted feature sets represent a part of a gesture with a high variability. For instance, NT57 covers different retraction movements and thus possesses no prototypical movement trajectory.

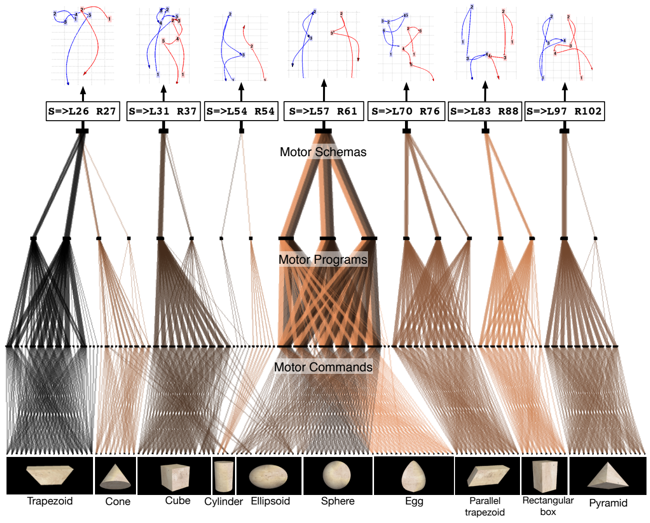


Figure 5: Resulting motor grammar as a hierarchical motor knowledge, where gestural movements with similar iconic mappings are grouped into the same schemas.

Figure 5 illustrates how the learned FSCFG can serve as a compositional/hierarchical representation of motor knowledge, which is learned from observed gesture performances. For this purpose, we consider each start rule equivalent to a motor schema and each motor program is represented through a pair of hand-specific rules (or a single one in the case of one-handed performances). Each motor command is represented through a terminal (or equivalently its producing lexical nonterminal). In Figure 5, the nodes at the bottom represent the individual gestures, with performances for different objects shown in different colors. Each of the outgoing connections from each gesture represents one of its movement segments which is then associated with the motor command, whose lexical rule is invoked for parsing that segment. Each outgoing connection, from commands to programs up to the start rules as schemas, represents a gesture performance.

As shown, gestures for the same or similar objects get associated with the same start rules. For example, since drawing gestures depict a rough silhouette of the referred object and subtle distinctions are ignored, gestures for sphere, ellipsoid and egg are clustered into a single start rule for "round" objects. This automatic generalization can be viewed as capturing the essence of depicting the corresponding iconic mapping (e.g. a general motor schema for depicting round shapes). Likewise, the gestures for cone and trapezoid are also parsed by the same start rule, but then get differentiated at the motor program level as different performances.

Finally, to demonstrate the important ability of FSCFG to also produce new gesture performances, the top row of Figure 5 shows a generated prototypical gesture (as wrist trajectories) for each motor schema. This shows that the proposed FSCFG cannot only be used to learn compositional motor knowledge at different levels of abstraction from highly varied gesture performances, but it can also be used to generate prototypes of the learned gestures, e.g. in a humanoid virtual agent or robot. Furthermore, learned FSCFGs can also be applied as discriminative models to recognize hand gestures, as we presented and discussed in (Sadeghipour & Kopp, 2014).

Conclusion

The goal of the present work is to investigate the structures of iconic gesturing at a level of movement features that affords a motor grammar. To deal with the lack of clear-cut symbolic units in gesture, we proposed the framework of FSCFG, which combines feature-based representation (incl. symbolization) with syntactical rule-based organization. That is, grammar-like structure is built on local, statistically identified primitives. Applied to natural iconic gesture trajectories, the resulting grammar organizes motor knowledge hierarchically across the levels of motor commands, programs, schemas. The well-known phasal organization of gesture is reified and, crucially, structures become visible that may fulfill independent depictive functions. This elevates motor knowledge to a level of structural organization, where significant invariant features of iconic mapping are identified that can provide first hints for a motor-meaning interface in communicative bodily behavior.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in 'Cognitive Interaction Technology'.

References

Chen, Q., Georganas, N. D., & Petriu, E. (2008). Hand gesture recognition using haar-like features and a stochastic context-free grammar. *Instrumentation and Measurement, IEEE Transactions on*, *57*(8), 1562–1571.

Flash, T., & Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology*, *15*(6), 660–666.

Guerra-Filho, G., & Aloimonos, Y. (2007). A language for human action. *Computer*, *40*(5), 42–51.

Hamilton, A., & Grafton, S. (2007). The motor hierarchy: From kinematics to goals and intentions. In *Attention and performance.* Oxford University Press.

Hong, P., Turk, M., & Huang, T. S. (2000). Gesture modeling and recognition using finite state machines. In *Automatic face and gesture recognition, 2000. proceedings. fourth ieee international conference on* (pp. 410–415).

Ivanov, Y., & Bobick, A. (2000). Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(8), 852–872.

Kendon, A. (1972). Some relationships between body motion and speech. In A. Siegman & B. Pope (Eds.), *Studies in dyadic communication.* New York: Pergamon Press.

Kitani, K. M., Sato, Y., & Sugimoto, A. (2006). *An mdl approach to learning activity grammars* (Vol. 106; Technical Report No. 376). Tokyo, Japan: IEICE - The Institute of Electronics, Information and Communication Engineers.

Kopp, S., Bergmann, K., & Kahl, S. (2013). A spreading-activation model of the semantic coordination of speech and gesture. In (pp. 823–828). Cognitive Science Society.

McNeill, D. (2000). *Language and gesture* (Vol. 2). Cambridge University Press.

Minnen, D., Essa, I., & Starner, T. (2003). Expectation grammars: leveraging high-level expectations for activity recognition. In *Computer vision and pattern recognition, ieee conference on,* (Vol. 2, pp. 626–632).

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, *37*(3), 311-324.

Montgomery, K. J., Isenberg, N., & Haxby, J. V. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive and Affective Neuroscience*, *2*(2), 114-122.

Mussa-Ivaldi, F. A., & Bizzi, E. (2000). Motor learning through the combination of primitives. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *355*(1404), 1755–1769.

Mussa-Ivaldi, F. A., & Solla, S. A. (2004). Neural primitives for motion control. *IEEE Journal of Oceanic Engineering*, *29*(3), 640-650.

Pastra, K., & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical transactions of the Royal Society of London. Biological sciences*, *367*(1585), 103–117.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, *11*(2), 416–431.

Ryoo, M. S., & Aggarwal, J. (2006). Recognition of composite human activities through context-free grammar based representation. In *Computer vision and pattern recognition, 2006 ieee computer society conference on* (Vol. 2, pp. 1709–1718).

Sadeghipour, A., & Kopp, S. (2011). Embodied gesture processing: Motor-based integration of perception and action in social artificial agents. *Cognitive Computation*, *3*, 419–435.

Sadeghipour, A., & Kopp, S. (2014). A hybrid grammar-based approach for learning and recognizing natural hand gestures. In *Proceedings of the 28th AAAI conference on artificial intelligence.* (in press).

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Unpublished doctoral dissertation, University of California at Berkeley, Berkeley, CA.

Tsai, W. H., & Fu, K. S. (1980). Attributed grammar - a tool for combining syntactic and statistical approaches to pattern-recognition. *IEEE Transactions on Systems Man and Cybernetics*, *10*(12), 873–885.

Zhang, Z., Tan, T., & Huang, K. (2011). An extended grammar system for learning and recognizing complex visual events. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(2), 240–255.