Modeling Action Verb Semantics Using Motion Tracking

Timo Honkela[1] and Klaus Förger[2]
[1]Department of Information and Computer Science
[2]Department of Media Technology
Aalto University School of Science, FI-00076 Aalto, Finland

Abstract

In this article, we consider how semantics of action verbs can be grounded on motion tracking data. We present the basic principles and requirements for grounding of verbs through case studies related to human movement. The data includes high-dimensional movement patterns and linguistic expressions that people have used to name these movements. We discuss open issues and possibilities related to symbol grounding. As a conclusion, we find the grounding to be useful when reasoning about the meaning of words and relationships between them within one language and potentially also between languages.

1 Introduction

The basic scientific question behind this article is how to computationally model the interrelated processes of interpreting natural language and perceiving movement in multimodal real world contexts. Namely, an important problem in natural language processing and in computer science in general is that in most cases computer systems processing symbols or language do not have access to the phenomena being referred to. In contrast, humans can readily associate expressions with their non-linguistic experiences and actions in the world. For instance, we know the different interpretations of color red in expressions "red skirt", "red skin" and "red wine" or the phrase "a long jump" may refer to very different things depending on the context. As a direct consequence, computational systems can only reason about the symbols themselves rather than about the grounded meaning or external references of those symbols. However, if we want machines to learn and use language as it is actually used by humans, we have to take into account that language is fully understood only through its use in linguistic and multimodal contexts [1].

In this article, we consider a seemingly simple domain of symbol grounding, naming human movement. It is, however, complex enough to be a non-trivial case which is also illustrated by the fact that different languages divide the space of body-related expressions in different ways [2]. Moreover, people may have different interpretations even regarding what they call "running", "jogging" or "walking" in less prototypical cases. Studying these differences is enabled by having access to the actual patterns of movement.

Extracting semantic information from the statistical regularities in large text corpora is nowadays commonplace. One obvious reason for using a statistical approach is cost-effectiveness: models of language can be built with less human effort than when traditional means are used. While statistical analysis of word occurrences and other linguistic constructions in their textual contexts has proven to be useful, there is a limit to how much can be inferred from texts only, and therefore obtaining data of words in their multimodal contexts is an important research topic. This kind of external contextualization is often referred to as symbol grounding [3].

Research on symbol grounding is multidisciplinary and multifaceted. Related to motion tracking, successful systems that achieve good classification results include identification of movement types [4], manners of moving [5] and the gender from walking movements [6]. The issue is, of course, relevant in robotics [7, 8]. In cognitive science, symbol grounding and embodiment is an important theme (cf., e.g., [9, 10, 11, 12, 13]). In a classical work, Bailey developed a computational model of the role of motor control in the acquisition of action verbs [14].

We are aware of the breadth and depth of the underlying philosophical [15] and methodological issues. In this article, we wish to address naming human movements as a concrete, limited but non-trivial case related to multimodally grounded natural language processing. In order to study how people name different human movements, we have used motion tracking to obtain data in which skeletons move on the screen. Using this data, we conducted two case studies. In the first study, we asked people to classify movements to a limited number of categories. The results of this classification task, serving as a feasibility study, are reported in the next section.

In the second case study, we asked people to describe these movements with their own words. It was important that the question was open ended because we wished to study the naming of movements which is different from classification. In naming, the labels given typically follow a Zipfian distribution [16]. The results of this case study are reported discussed in Section 3.

2 Grounding through motion capture

The motion tracking has been conducted using OptiTrack Motion Capture system and the ARENA software, developed by NaturalPoint, Inc. We recorded 16 minutes of human motion which was manually annotated with the following labels: jump, sit down, sitting, stand up, standing, turn left, turn right, walking and waving hand. The labels were allowed to overlap as for example walking and waving hand can be done at the same time.

Four types of features were extracted from the data (see Figure 1). The first type was absolute values of velocities of all the body parts. The second type was distances between the end parts of the limbs. The third was velocity vectors of the end parts of the limbs. The last type was coordinate positions of the end parts of the limbs. To make the velocity vectors and positions usable, we had to center the coordinate system to the hips of the character and rotate the character to always face the same direction. This resulted in 72 feature dimensions in the first case study. We averaged the values of the features over 0.25 seconds to get the final values used in the classification.

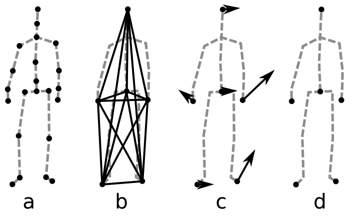

Figure 1: The types of the features used in the classification include absolute velocities for each body part (a), distances between limb ends (b), velocity vectors of limb ends (c), and positions of limb ends (d).

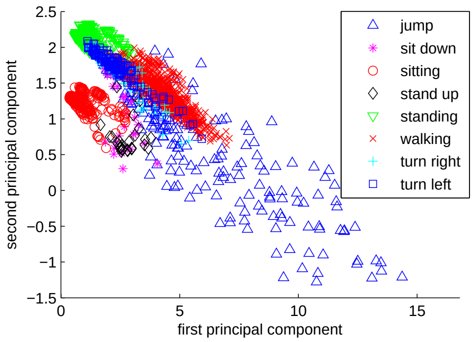

Figure 2: Training samples used the classification plotted along the first two principal components.

Grounding verbs requires associating them with patterns of motion based features. A good way to ensure that the used features are not only random numbers is to see how well the features can be used in classifying previously unseen motions. To classify the data we used K nearest neighbors with a Euclidean distance metric. The classification was tested on two minutes of motion that was not used in the training set with results at the same level as obtained earlier by others [4, 5, 6]. In classification, the transition motions between two verbs were the main problem. The classifier tried to forcible classify the motion when the most natural option would be not giving a class at all as the transitions may not correspond to any verb. One reason for the good performance lies in the well-selected features. When the training samples are plotted along the first two principal components (see Figure 2), it becomes evident that many of the classes are separated by the used features.

The features form a space where all individual frames of a motion can be projected. As two consecutive frames of motion are always similar due to physical

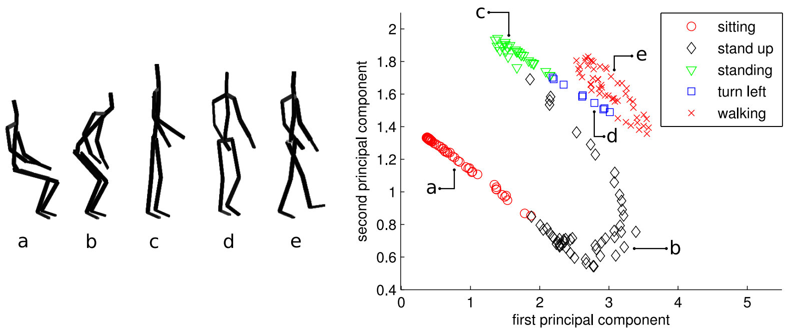

Figure 3: On the left hand side, motion of a character (a) sitting, (b) standing up, (c) standing, (d) turning left and (e) walking (left), and on the right hand side, the trajectory formed by the frames plotted on the first and second principal component.

restrictions, motions can be plotted as trajectories in the feature space. This is shown in Figure 3, where a motion starting from sitting, going through standing up, standing and walking, is plotted along the first two principal components of the feature space.

The separation by the two principal components is not complete as the data is inherently high dimensional as can be seen in Figure 4. The figure shows that more than 10 principal components are needed in order to explain 90% of the variance in the data.

The fact that many labels can be valid for a motion simultaneously is a challenge for using the features of the classification as distance measures. For example, waving hand can be done while walking, sitting or standing. This is visible in the training samples used for those classes in Figure 5. As 'waving hand' appears in several separate clusters, the mean distance between it and other labels does not reflect the real relations between the labels. Therefore, the overlap between labels should be analyzed before similarity of the labels.

3 Modeling relations between verbs

In order to have a fine-grained collection of movements, we asked actors to perform walking, limping and running in various styles. These movements were blended in three steps including alignment, time warping and interpolation. In time warping, the motions were synchronized. In the third step, the coordinates of the root joints were interpolated linearly and the joint rotations were interpolated in a four-

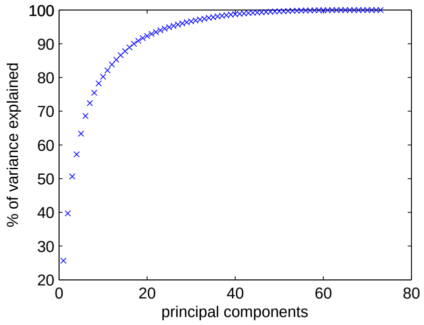

Figure 4: Variance explained by the principal components plotted cumulatively.

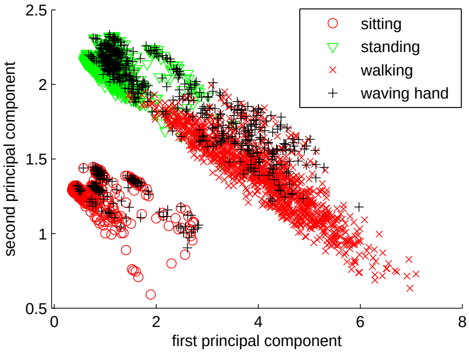

Figure 5: Training samples with label 'waving hand' and other samples with labels that can be used simultaneously plotted along the first two principal components.

coordinate system [17]. The blending enabled us to have a larger number of variations of the movements. The people were asked to describe the movement with one verb or phrase. The task was to label 24 to 124 videos where the videos lasted from 3 to 12 seconds. Each video was portraying a stick figure representation.

We analyzed the questionnaire results where 22 persons had named the movements in Finnish language. We used the self-organizing map (SOM) [18] algorithm to conduct a mapping from the 602-dimensional movement space into a 2-dimensional display. The movement determines the map structure and structure of labels is obtained by including them in the input vector with a small weight. To illustrate the outcome, we chose 12 verbs to be analyzed in more detail. This map of labels is shown Fig 6. With one exception, each verb is associated with a contiguous area on the map. For instance, the verb “walk” is located on the upper side of the map and the verb “run” on the lower left corner. Fig 7 shows examples of underlying movement features that have determined the organization of the map. The area for running in Fig 6 coincides with the feature “mean acceleration of hips” in Fig 7. The union “running” and “jogging” coincides with the distribution of features “mean absolute velocity of hips” and “mean absolute velocity of abdomen”.

In many cases, the association between the labels and patterns of movement is not one-to-one but require consideration of a reasonably large number of features. On the other hand, the vector space for the associations is much lower in dimensionality than the pixel patterns over time in the original videos. This is thanks

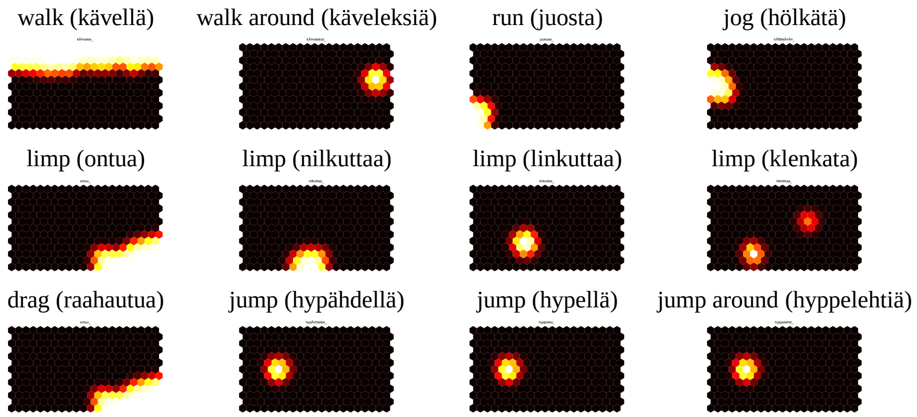

Figure 6: The distribution of 12 verbs on a self-organizing map of movements. A light shade denotes a high value of each verb.

to the motion tracking system that compresses the original very high dimensional feature space into a large number of meaningful variables. In the general case, it remains a challenge how to conduct the pattern recognition and dimensionality reduction in such a way that relevant features are included for the associations. In many early studies the low-level representations were based on manually encoded structures (cf., e.g., [14]). In order to develop large scale solutions, the process should, however, be as automatic as possible. Due to variety of applications that may require different kinds of feature sets for same domain, the features extraction process needs to be task-dependent [19].

4 Conclusions and discussion

We are interested in answering the scientific question of how to enable machines to have a increasingly common ground with humans for associating language with perceptual patterns. In the following, we discuss two symbol grounding themes.

4.1 Multimodally grounded language technology

Through multimodally grounded language technology, more robust and correct manipulation of linguistic data becomes possible, e.g., when resolving ambiguities or when needing deeper inference. Application areas include building animations using linguistic instructions and coaching of skills.

What centrally constrains communication is the dissimilarity of the conceptual systems of the discussants. An important aspect of better understanding of human

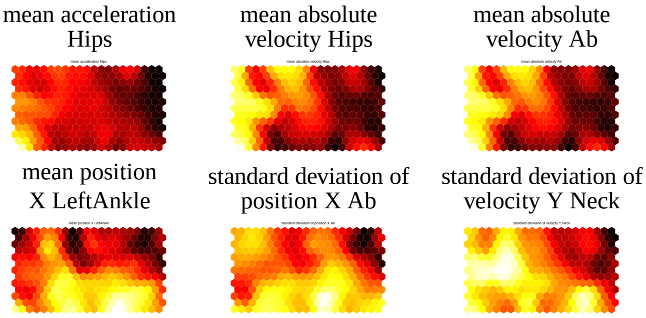

Figure 7: The distribution of 6 features out of 602 on the self-organizing map of movements. A light shade denotes a high value of each feature.

expression is, we believe, capturing human conceptualizations of the environment in which the co-operation is to take place. The subjective aspect of interpretation can be analyzed when the use of symbols is considered in different contexts by a number of individuals [20].

4.2 Multimodally grounded translation

It has earlier been demonstrated that the association with visual information can be used even to find parallels between different languages [21]. An analysis of the similarities in the visual appearance of some object can be used to find a conceptual link between a word in one and another language. This is analogical to ostensive definition, based on pointing out examples. In the future, we plan to collect labeled data in multiple languages. This enables developing a mapping function between action verbs in different languages based on the common ground.

Acknowledgments

We gratefully acknowledge the financial support by Academy of Finland for the Multimodally Grounded Language Technology project (254104) and HeCSE graduate school that have made this research possible. We also wish to thank our colleagues in the project who have helped in conducting the research in various ways including Prof. Tapio Takala, Dr. Jorma Laaksonen, Dr. Markus Koskela, Dr. Harri Valpola, Dr. K. Lagus, Ms. Xi Chen, Mr. Paul Wagner and Mr. Oskar Kohonen. The insights in this area have also progressed thanks to collaborations with Prof. Peter Gärdenfors (Lund University, Sweden) and Prof. Lars Kai Hansen (DTU, Denmark) and their teams.

References

[1] Hörmann, H.: Meaning and Context. Plenum Press, New York (1986)

[2] Choi, S., Bowerman, M.: Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. Cognition **41**(1) (1991) 83–121

[3] Harnad, S.: The symbol grounding problem. Physica D **42** (1990) 335–346

[4] Pavlović, V., Rehg, J., MacCormick, J.: Learning switching linear models of human motion. In: Advances in Neural Information Processing Systems 13. (2001) 981–987

[5] G.W. Taylor, G.E. Hinton, S.R.: Modeling human motion using binary latent variables. In: Advances in Neural Information Processing Systems 19. (2007) 1345–1352

[6] Davis, J., Gao, H.: Gender recognition from walking movements using adaptive three-mode PCA. In: Computer Vision and Pattern Recognition Workshop. (2004)

[7] Roy, D.: Grounding words in perception and action: computational insights. Trends in cognitive sciences **9**(8) (2005) 389–396

[8] Williams, M.A., McCarthy, J., Gärdenfors, P., Stanton, C., Karol, A.: A grounding framework. Autonomous Agents and Multi-Agent Systems **19**(3) (2009) 272–296

[9] Lakoff, G., Johnson, M.: Philosophy in the Flesh - The Embodied Mind and its Challenge to Western Thought. New York: John Wiley (1999)

[10] Glenberg, A.M., Robertson, D.A.: Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of memory and language **43**(3) (2000) 379–401

[11] Sun, R.: Symbol grounding: a new look at an old idea. Philosophical Psychology **13**(2) (2000) 149–172

[12] Vogt, P.: The physical symbol grounding problem. Cognitive Systems Research **3**(3) (2002) 429–457

[13] Gärdenfors, P., Warglien, M.: Using conceptual spaces to model actions and events. Journal of semantics **29**(4) (2012) 487–519

[14] Bailey, D.: When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs. PhD thesis, UC Berkeley (1997)

[15] Honkela, T.: Philosophical aspects of neural, probabilistic and fuzzy modeling of language use and translation. In: Proceedings of IJCNN 2007. (2007) 2881–2886

[16] Li, W.: Zipf's law everywhere. Glottometrics **5** (2002) 14–21

[17] Shoemake, K.: Animating rotation with quaternion curves. SIGGRAPH Computer Graphics **19**(3) (1985) 245–254

[18] Kohonen, T.: Self-Organizing Maps. Springer (2001)

[19] Ji, R., Yao, H., Liu, W., Sun, X., Tian, Q.: Task-dependent visual-codebook compression. Image Processing, IEEE Transactions on **21**(4) (2012) 2282–2293

[20] Honkela, T., Raitio, J., Nieminen, I., Lagus, K., Honkela, N., Pantzar, M.: Using GICA method to quantify epistemological subjectivity. In: Proc. of IJCNN 2012. (2012) 2875–2883

[21] Sjöberg, M., Viitaniemi, V., Laaksonen, J., Honkela, T.: Analysis of semantic information available in an image collection augmented with auxiliary data. In: Proc. of AIAI'06, Artificial Intelligence Applications and Innovations, Springer (2006) 600–608