Learning Semantics of Movement

Timo Honkela[1] **Oskar Kohonen**[1] **Jorma Laaksonen**[1]

Krista Lagus[1] **Klaus Förger**[2] **Mats Sjöberg**[1]

Tapio Takala[2] **Harri Valpola**[1,3] **Paul Wagner**[1]

[1] Aalto University School of Science, Department of Information and Computer Science
[2] Aalto University, School of Science, Department of Media Technology, Espoo, Finland
[3] ZenRobotics Ltd., Helsinki, Finland

1 Extended abstract

The basic scientific question behind this presentation is how to model computationally the interrelated processes of understanding natural language and perceiving and producing movement in real world contexts. Rather than relying on manually defined representations, we approach this problem by analyzing the statistical regularities in contexts of language use. A central aspect is that the learning of internal representations is based on large amounts of fully or mostly unlabeled training data [1, 2]. A promising methodological approach towards this is that data from different sources can be brought together through multiview/transfer/multi-task learning (see e.g. [2, 3, 4, 5]). Another promising approach is learning relations. Relations are powerful abstractions because they allow reasoning not just about objects, but about their combinations. Recent work suggests that at least limited classes of relations can be learned from data [6]. We have earlier conducted research in which text contexts have been used to learn semantic similarities (see e.g. [7, 8]. However, to reach a more human-level understanding, we have to take into account that language is fully understood only through its use in its multimodal and embodied contexts including linguistic, visual, auditory, tactile and kinestetic dimensions. In general, patterns and signals are natural representations for multimodal contexts. These differ considerably from the discrete representations of symbols and expressions in symbolic languages. For instance, images are typically represented as numerical matrices. We have earlier conducted research related to the combination of image and language data (see e.g. [9, 10]).

Movement is the specific focus of this presentation for several reasons. It is a fundamental part of human activities that ground our understanding of the world. For a computational agent, information about movement is important to its world knowledge because movement underlies an essential part of the semantics of human languages. Movement is so central to the human embodiment that it is used as a basis for understanding both concrete and abstract utterances [11]. Abstract meanings are often constructed as metaphoric extensions of movement schemas. Furthermore, movement seems to be a less studied topic in learning semantics from multimodal data. As there is an increasing amount of video and motion tracking data available, formation of semantic models based on movement using computational methods is becoming feasible. Applications of movement-based learning of semantics include animation [12], sports instruction [13], context-sensitive reasoning in ubicomp services and improved disambiguation in natural language processing tasks.

In cognitive linguistics, it is argued that the meaning of linguistic symbols are representations in the mind of the language users that derive from the users' sensory perceptions, their actions with the world and with each other. For example, the meaning of the word 'walk' involves what walk-

ing looks like, what it feels like to walk, and after having walked, what the world looks like while walking (e.g. objects approach at a certain speed). In certain domains it is possible to measure information that can be used to form a model of the phenomenon similar to the one that is hypothesized to exist in the human brain. In particular, it appears that the representation of space, of movement in space and actions among objects in space underlies much of linguistic cognition, and therefore affects the meanings of words and expressions, and the way we generate and understand language [14, 15, 16]. Research within cognitive linguistics has shown that one basic mechanism in language is "metaphorical generalization" from concrete human 3D movement and spatial relations to abstract concepts.

In a project that started recently, we develop methods and technologies to automatically associate human movements detected by motion capture and in video sequences with their linguistic descriptions. When the association between human movement and their linguistic descriptions has been learned using pattern recognition and statistical machine learning methods, the system is also used to produce animations based on written instructions and for labeling motion capture and video sequences. This setting lends itself to multi-task learning. The link between movement and language is also examined in relation to the context and the quality and nature of the movement. A secondary objective is to create a library of movements with their corresponding labels that can be used for the development and training of the machine learning methods.

Motion tracking data captures the most important properties of movement directly, whereas extracting them from video data is nontrivial. The motion tracking system can capture all types of human body movement including both subtle and fast motion. Skeletal movement can be tracked as long as a sufficient number of markers are visible. Also existing motion capture data sets that are available will be examined and may be utilized (see e.g. [17]). There are examples of successful classification systems that achieve good classification results for related problems, including identifying movement types ([18, 19]), manners of moving ([20]) and the gender from walking movements ([21]). [22] have been able to distinguish active emotions (joy, anger) from passive emotions (relief, sadness) based on 2D motion extracted from videos. We have similar observations with captured motion of walking [23] and conducting music [24]. Our current work on video analysis is based on the existing PicSOM content-based multimedia analysis, search and retrieval system, successfully applied in numerous international evaluations in image and video categorization and search [25]. The system supports a large variety of sub-methods necessary for analyzing video streams from different sources. Such components include standard low-level descriptors for visual data, face detection and recognition, combining evidence from multiple sources and modalities, and various machine learning algorithms. In our recent projects, the system has been applied, e.g., to motion analysis for automatic video summarization, to generic semantic concept detection from video material [26] and also to the analysis of Sign Language video streams. When the video and motion capture recordings are done simultaneously, one can readily compare analyses of those two synchronous data sources and use the motion capture data as a ground truth for the video-based analysis. Animation requires generation of movement, which can also be beneficial for spatial reasoning. There are several ways to generate movement, including using motions from a database, interpolation of motions [27], Switching Linear Dynamic Systems [18], Bayesian methods [28] and Deep Belief Networks [20], of which the last one can also be trained for the manner of movement, such as a particular walking style.

For learning relations between objects we have recently developed a method for identifying pairwise relations between moving objects [6]. The method is related to Denoising Source Separation (DSS) [29] and is a fast method for discovering whether a mapping (relation) holds between data sets. It is applied to feature lists of segmented objects, and discovers changing relations between the objects (e.g. object A follows object B). Applying this kind of method to the movement data is something we will do in the future.

In our presentation, we review earlier results and discuss new directions for learning semantics of movement. We consider four different aspects: using video and motion tracking data, applying multi-task learning methods, learning relations, and framing the problem within cognitive linguistics research. Moreover, it appears that analyzing movement can have an important role in learning semantics in general, because of the central role movement plays in language.

Acknowledgments

We gratefully acknowledge Academy of Finland's support through "Multimodally Grounded Language Technology" project and Centre of Excellence in Adaptive Informatics Research.

References

[1] T. Honkela, A. Hyvärinen, and J. Väyrynen. Wordica - emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering,*, 16(3):277–308, 2010.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, (to appear).

[3] A. Bordes, N. Usunier, and J. Weston. Label ranking under ambiguous supervision for learning semantic correspondences. In *Proc. of ICML'2010*, pages 103–110, 2010.

[4] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Proc. of the 10th Eur. Conf. on Computer Vision*, 2008.

[5] G. Leen, J. Peltonen, and S. Kaski. Focused multi-task learning using gaussian processes. In *Proceedings of the ECML/PKDD 2011*, 2011.

[6] O. Kohonen, H. Valpola, and K. Lagus. Learning to detect roles and relations. (manuscript).

[7] T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in grimm tales, analyzed by self-organizing map. In *Proc. of ICANN'95, vol. II,*, pages 3–7, 1995.

[8] K. Lagus, A. Airola, and M. Creutz. Data analysis of conceptual similarities of finnish verbs. In *Proc. of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society*, pages 566–571, 2002.

[9] M. Sjöberg, V. Viitaniemi, J. Laaksonen, and T. Honkela. Analysis of semantic information available in an image collection augmented with auxiliary data. In *Proc. of AIAI'06, Artificial Intelligence Applications and Innovations, vol. 204*, pages 600–608. Springer, 2006.

[10] M. Sjöberg, J. Laaksonen, T. Honkela, and M. Pöllä. Inferring semantics from textual information in multimedia retrieval. *Neurocomputing*, 71(13–15):2576–2586, 2008.

[11] George Lakoff and Mark Johnson. *Philosophy in the Flesh - The Embodied Mind and its Challenge to Western Thought*. New York: John Wiley, 1999.

[12] S. Levine, C.Theobalt, and V.Koltun. Body language animation synthesis from prosody. *ACM Transactions on Graphics*, 28(5), 2009.

[13] H. Ghasemzadeh, V. Loseu, E. Guenterberg, and R. Jafari. Sport training using body sensor networks: a statistical approach to measure wrist rotation for golf swing. In *Proceedings of the Fourth International Conference on Body Area Networks*, BodyNets '09, pages 2:1–2:8, ICST, Brussels, Belgium, 2009. ICST.

[14] M. Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago, 1990.

[15] T. Regier. *The Human Semantic Potential: Spatial Language and Constrained Connectionism.* MIT Press, Cambridge, 1996.

[16] D. Bailey. *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.

[17] Y.L. Ma, H. Paterson, and F.E. Pollick. A motion-capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods*, 38(1):134–141, 2006.

[18] V. Pavlović, J.M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems 13*, pages 981–987, 2001.

[19] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984–989, 2005.

[20] S.T. Roweis G.W. Taylor, G.E. Hinton. Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems 19*, pages 1345–1352, 2007.

[21] J.W. Davis and H. Gao. Gender recognition from walking movements using adaptive three-mode pca. In *Computer Vision and Pattern Recognition Workshop*, 2004.

[22] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *Computer Vision and Pattern Recognition Workshops IEEE*, pages 1–6, 2008.

[23] K. Lehtonen and T. Takala. Evaluating emotional content of acted and algorithmically modified motions. In *24th Annual Conference on Computer Animation and Social Agents (CASA2011)*, Lecture Notes in Computer Science 6758, pages 144–153. Springer, 2011.

[24] T. Ilmonen and T. Takala. Detecting emotional content from the motion of an orchestra conductor. In Courty Gibet and Kamp, editors, *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop GW 2005*, Lecture Notes in Computer Science 3881, pages 292–295. Springer, 2006.

[25] V. Viitaniemi and J. Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, 22(6):557–568, 2007.

[26] M. Sjöberg, M. Koskela, M. Chechev, and J. Laaksonen. Picsom experiments in trecvid 2010. In *Proc. of the TRECVID 2010 Workshop*, Gaithersburg, MD, USA, 2010.

[27] O. Arikan, D. A. Forsyth, and J. F. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics*, 22(3):402–408, 2003.

[28] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems 22*, pages 549–557, 2009.

[29] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.