Journal of Ambient Intelligence and Smart Environments 1 (2009) 15–21
DOI 10.3233/AIS-2009-0002
IOS Press

Sensory grammars for sensor networks

Yiannis Aloimonos
Computer Vision Laboratory, Institute for Advanced Computer Studies, Computer Science Department, Cognitive Science Program, University of Maryland, College Park, MD 20742, USA. E-mail: yiannis@cs.umd.edu

Abstract. One of the major goals of Ambient Intelligence and Smart Environments is to interpret human activity sensed by a variety of sensors. In order to develop useful technologies and a subsequent industry around smart environments, we need to proceed in a principled manner. This paper suggests that human activity can be expressed in a language. This is a special language with its own phonemes, its own morphemes (words) and its own syntax and it can be learned using machine learning techniques applied to gargantuan amounts of data collected by sensor networks. Developing such languages will create bridges between Ambient Intelligence and other disciplines. It will also provide a hierarchical structure that can lead to a successful industry.

Keywords: Sensor networks, human action, human activity languages, sensorimotor representations

1. Introduction: Sensor networks

The field of ambient intelligence and smart environments has been flourishing over the past several years. For different reasons, governments and industry showed a major interest in further developing this technology. At the same time, the problems surrounding this field represent interesting problems for modern computer science that is concerned with gargantuan amounts of data. From a basic viewpoint, this discipline amounts to developing sensor networks that "sense" their environment and take appropriate actions [2–4].

Successes in sensor networks and the infrastructure they have produced are leading people to consider the deployment of sensor networks in everyday life situations such as assisted living, workplace safety and entertainment [8,20]. These new applications however induce a different operation paradigm. Instead of focusing on the collection of raw data to be analyzed by domain experts, this new breed of sensor networks is expected to summarize and interpret raw sensor data into a set of higher level semantics that will serve as the building blocks for providing a variety of services. Consider everyday life environments with an abundance of sensors. Some of them are part of the infrastructure such as cameras on the ceiling and walls, RFIDs on objects and Zigbee sensors on appliances. Others are carried by people (e.g. GPS and cameras on cell-phones), and some of them are deployed for a specific application. All these sensors can provide a lot of information, but there is still a major challenge. How does one yield the power of such networks? How can the network discern and interpret the useful information from a heterogeneous set of sensor measurements?

Considering the commonalities between the types of data interpretation needed in everyday life applications of sensor networks, one can easily realize that the network needs to interpret data according to a set of rules and patterns. In an elder care situation for instance, the doctors or the primary care givers would use an in-home sensor network to monitor activity levels, detect "falls" or getting "stuck" situations, and to prevent people from engaging in risky behaviors. In a workplace safety application, the supervisor would most likely want to task a sensor network to check for a set of safety conditions in order to mitigate risks and liabilities. In more abstract applications, one could check the state of a system against a specification of how the system should operate to identify when faults take place. This could be applied on the sensor network itself to detect faults or to check for security violations.

How are we going to achieve this? Every time we have a new application, we must come up with the rules and patterns? This is what is actually happening today in the state of the art. In the recent literature, one will note a number of very interesting approaches for interpreting the sensory data stream.

Many of them are based on very simple features that can be extracted from the data, yet such features could be sufficient for solving a specific problem. An interesting example consists of finding the amount of time a person spends at a particular location. Due to the vast amount of training data, the appropriate distributions of timing could be built and used for inference [17,18,31].

Looking into the future, however, it is clear that we must adopt a more basic approach. After all, sensor networks "sense" interacting and behaving humans. In other words, the sensor networks of the future will need to interpret human activity. In the sequel, a research program along this line of thought is outlined, assuming that the sensors are ordinary video cameras.

2. Human action: It resides in many spaces

One of the important lessons from the field of the Neurosciences [7,10,16,24] is the model of action shown in Fig. 1. Before the command is sent, a copy is kept (the efference copy). The efference copy can be used with forward models and predicted feedback in order to "think" about an action, without actually doing it. In other words, we have inside our minds abstract representations of actions, our own and of others. It is those representations that sensor networks of the future should be extracting.

Knowledge of actions is crucial to our survival. Hence, human infants begin to acquire actions by watching and imitating the actions performed by others. With time, they learn to combine and chain simple actions to form more complex actions. This process can be likened to speech, where we combine simple constituents called phonemes into words, and words into clauses and sentences. The analogy does not end here: humans can recognize as well as generate both actions and speech. In fact, the binding between the recognitive and generative aspects of actions is revealed at the neural level in the monkey brain by the presence of mirror neuron networks, i.e., neuron assemblies which fire when a monkey observes an action (like grasping), and also when the monkey performs the same action [10]. All these observations lead us to a simple hypothesis:

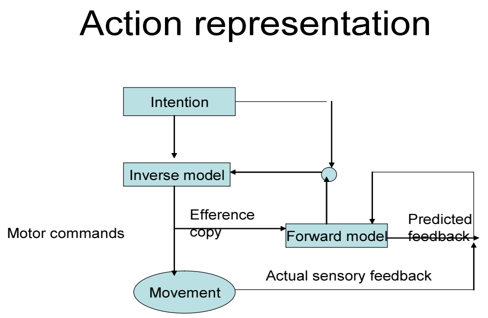

Fig. 1. Contemporary model of human action.

Actions are effectively characterized by a language. This is a language with its own building blocks (phonemes), its own words (lexicon) and its own syntax.

The realm of human actions (e.g., running, walking, lifting, pushing) may be represented in at least three domains: visual, motor, and linguistic. The visual domain covers the form of human actions when visually observed. The motor domain covers the underlying control sequences that lead to observed movements. The linguistic domain covers symbolic descriptions of actions (natural language – English, French, etc.). Thus, it makes sense to take the hierarchical structure of natural language (e.g., phonology, morphology, syntax) as a template for structuring not only the linguistic system that describes actions, but also the visual and motor systems. One can define and computationally model visual and motor control structures that are analogous to basic linguistic counterparts: phonemes (the alphabet), morphemes (the dictionary), and syntax (the rules of combination of entries in the dictionary) using data-driven techniques grounded in actual human movement data. Cross-domain relations can also be modeled, yielding a computational model that grounds natural language descriptions of human action in visual and motor control models. Since actions have a visual, motor and a natural language, converting from one space to another becomes a language translation problem (Fig. 2).

Thus we should be after a methodology for grounding the meaning of actions, ranging from simple movement to intentional action (e.g., go from A to B), by combining the (hypothesized) grammatical structure of action (motoric and visual), with the grammatical structure of planning or intentional action. This way, having an understanding of the grammar of this language, we should be able to parse the measurements from the sensors.

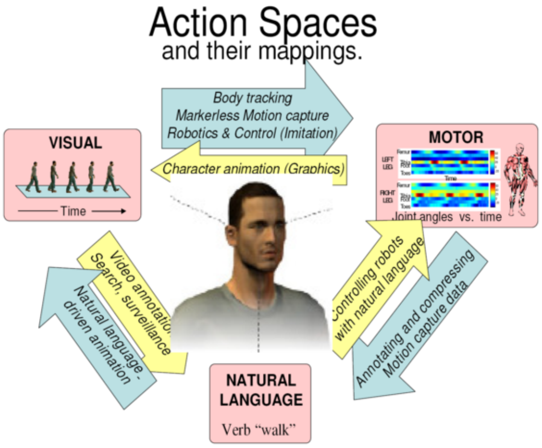

Fig. 2. Three action spaces (visual, motoric, natural language); many interesting problems in today's HCC,HCI&HRI become translation problems from one space to another, e.g. video annotation, natural language driven character animation, imitation, and so on.

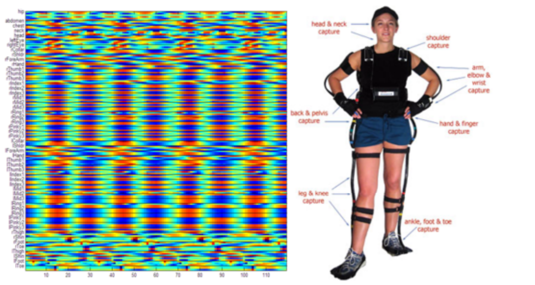

Fig. 3. Right: A motion capture suit providing data from human movement in a wireless manner. Motion capture data amounts to 3D trajectories of the joints. An equivalent representation actually captured by the suit showing a time evolution of joint angles is depicted on the left. For each joint (vertical axis) there are at most 3 varying rotation angles over time (horizontal axis). Each "row" is a 1D function (coded in shades of grey).

3. Languages of human action: They can be learned

By language we refer to what is conventional in modern computer science and computational linguistics, namely a system consisting of three interwined sub-systems: phonology, morphology and syntax [14]. Phonology is concerned with identifying the primitives (phonemes/letters) that make up all actions [*phonemic differences are often defined as the smallest change in the surface form (of speech) that signals a difference in meaning; a distinction that is phonemic in one language may not be phonemic in another*], morphology is concerned with the rules and mechanisms that put together the primitives into morphemes (words/basic actions) [*morphemes are the basic units of language that have stand alone meanings. For example "unwanted" has three morphemes: un-, want, and –ed; "cat" and "dog" have one morpheme each; "cats" has two morphemes (the –s means plural)*] and syntax is concerned with the mechanisms that put together the words (actions) into sentences (complex, composite actions/behavior). There may be a theoretically deeper reason for cross-modal similarities, with the analogy being real in humans, in the sense that there is a "grammar of thought" that is reflected in natural language but also structures other cognitive domains.

With regard to motoric actions, we do not need to go down at the level of neurons and muscles (motors and actuators) in order to create measurements. After all we are not interested in a neural theory of action, but in a framework that advances sensor network interpretation. Instead we can consider a higher level description, like the one provided by motion capture systems (see Fig. 3). Currently, such systems are moving into the real world in the form of suits one can wear beneath the clothing, allowing us to collect motoric data for thousands of actions in natural settings. This data can be imported into commercial animation/graphics packages such as POSER or MOTION BUILDER, to produce videos from any viewpoint of the motoric actions. In addition, we can acquire video data of the person wearing the motion capture suit and performing actions. As a result, we have for the first time access to a very large amount of data containing actions in a motoric space (joint angles vs time) and visual space (images). In the spirit of today's *zeitgheist (these are the years of hyper-empiricism)* we can apply techniques from statistics and learning in order to compress the information in our dataset. If we are able to represent all these actions efficiently then we will indeed have a language.

There are two main ways to go about it. One (a learning approach) would be to let the data decide, i.e. to obtain through grammatical induction techniques [21] a probabilistic grammar generating all the actions in an observation set. The other way (synthetic modeling approach) would be to impose an apriori model for the primitives (phonemes) and then through the appropriate learning or compression discover the morphological grammars as well as syntax.

It is by now clear that we should be studying not only the visual space of action – even though that is the input from our sensor networks – but the motoric space as well. The lesson from the Neurosciences suggests that when humans see and understand an action they do so through an internal act that captures the essence of the action, i.e. they map the observed action to their own "potential" movements. From the viewpoint of the sensor network engineer, the additional motor space is an unexpected benefit.

4. Grammars of visual and motoric human movement

4.1. Visual grammars

We believe that the right place to begin a discussion about actions and their recognition is to first ask the question: what do we really mean by actions? When humans speak of recognizing an action, they may be referring to a set of visually observable transitions of the human body such as 'raise right arm', or an abstract event such as 'a person entered the room'. While recognizing the former requires only visual knowledge about allowed transitions or movements of the human body, the latter requires much more than purely visual knowledge: it requires that we know about rooms and the fact that they can be 'entered into' and 'exited from', along with the relationships of these abstract linguistic verbs to lower level verbs having direct visual counterparts. Current work [23] deals with the automatic view-invariant recognition of low level visual verbs which only involve the human body. The visual verbs enforce the visual syntactic structure of human actions (allowed transitions of the body and viewpoint) without worrying about semantic descriptions. In [23] each training verb or action a is described by a short sequence of key pose pairs a = ((p1, p2), (p2, p3), ..., pk), where each pose pi belongs to P, where P is the complete set of observed (allowed) poses. Note that for every consecutive pair, the second pose in the earlier pair is the same as the first pose in the latter pair, since they correspond to the same time instant. This is because what we really observe in a video is a sequence of poses, not pose pairs. Hence, if we observe poses (p1, p2, p3, p4) in the video, then we build the corresponding pose pairs as ((p1, p2), (p2, p3), (p3, p4)). Each pose pi is represented implicitly by a family of silhouettes (images) observed in m different viewspoints, i.e. pi = (p1i, p2i, ..., pmi). The set of key poses and actions is directly obtained from multi-camera multi-person training data without manual intervention. A probabilistic context-free grammar (PCFG) is automatically constructed to encapsulate the knowledge about actions, their constituent poses, and view transitions. During recognition, the PCFG is used to find the most likely sequence of actions seen in a single viewpoint video. Thus, in this language the phonemes are multi-view poses of the human body and actions amount to transitions among them. It is desirable to obtain languages where the phonemes amount to poses of the body parts. This will bring the visual language on some form of equivalence with the motoric language.

4.2. Motoric grammars

Motoric languages still remain hidden, although considerable progress has been made with the language HAL [11–13] and the Behaviourscope Project [19]. The fundamental question is related to the phonemes or primitives of the language, called here kinetemes. As one can see in Fig. 3, motoric action data amounts to a set of 1D functions and for most actions there is a high degree of coordination among the different joints. This is evident in Fig. 3, where we can immediately see two sets of rows (of 1D functions) that "go together". These are the synergies [22,30]. The brain cannot generate a large number of independent control movements (that's why juggling is hard!). It generates a few, but those few are sufficient for generating any movement. The trick is that basically the same signal is sent to a group of joints, and for each joint it is modified using a number of parameters.

The basic problem for the years to come is the discovery (or invention) of the primitives in human movement [15,28,32]. Let us assume that these control symbols will be some form of a basic function, let's say a wavelet. Then, Fig. 4 explains how a grammar of wavelets could produce human action. A grammar generates control symbols (wavelets) that a control mechanism turns into a function which in turn can generate all the functions in the synergy, by changing a few parameters for each joint.

Ultimately, one would be interested in visuo motor representations. In terms of Fig. 2, we would be interested in developing a map from the visual space to the motoric space. Given that we can acquire visual data of someone wearing a motion capture suit, it becomes feasible to learn this map from a very large number of examples.

5. Ambient intelligence in the service of health

Ambient intelligence, equipped with a grammar of human action, becomes a very important tool in a variety of arenas, including Health. Movement, be-

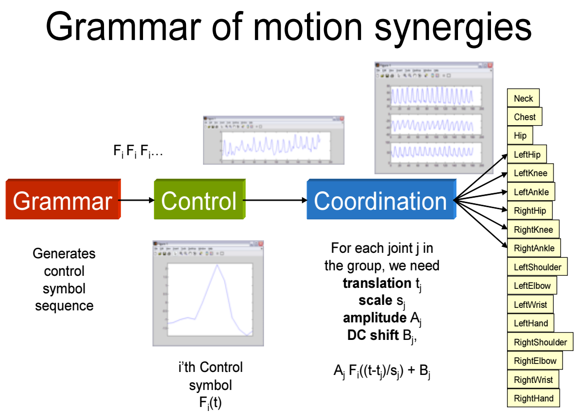
Grammar of motion synergies
F_i F_i F_i…
Grammar
Control
Coordination
Generates control symbol sequence
i'th Control symbol F_i(t)
For each joint j in the group, we need translation t_j scale s_j amplitude A_j DC shift B_j,
A_j F_i((t-t_j)/s_j) + B_j
Neck
Chest
Hip
LeftHip
LeftKnee
LeftAnkle
RightHip
RightKnee
RightAnkle
LeftShoulder
LeftElbow
LeftWrist
LeftHand
RightShoulder
RightElbow
RightWrist
RightHand

Fig. 4. Motion synergies are realized through wavelets and vectors of four parameters (translation, scale, amplitude and DC shift) for each joint. A coordination mechanism sends the wavelet to a group of joints where the vector of parameters modifies it before sending it for executing movement.

cause it is universal, easily detectable and possible to measure, has been a large window into the nervous system. Using, for the most part, measurements of human movement, Behavioral Neuroscience has had major accomplishments, such as documenting mile stones in human development and establishing a relationship between brain and behavior in typical and atypical populations. The measurements are per formed today with a cornucopia of sophisticated techniques, ranging from infrared and video to magnetic-based approaches, RFID and wireless sensor networks (with their advantages and disadvantages regarding accuracy, portability, intrusion, cost, etc.). In the future these measurements will be performed in "smart environments". However, despite the tremendous progress on measuring human movement, we still don't know, for example, how to track motor decline in elderly people during daily life activities at home and the workplace. With regard to Parkinson's disease, we still do not know how to assess, in quantitative terms, the effectiveness (tracking over time) of a new medicine. With regard to autism, we still do not know how social interaction deficits are manifested in body gestures, so that an early diagnosis can become possible. Why can't we yet deal with problems of such nature?

It is clear that the problems mentioned above have characteristics that are beyond the state of the art. To be able to track the evolution of Parkinson's disease it is not enough to just perform measurements of the human movement; we have to look at these measurements in a new, holistic sense. We must group the measurements into subsets that have "meaning" and then find global patterns and relationships in these groups. With regard to tracking motor decline, we need to interpret long sequences of measurements as interaction between two or more people. In the case of autism, we need to be able to pinpoint idiosyncratic characteristics of the whole body gesture or of a series of actions and interactions. In other words, we need to move to the next step, which is to study the structure of human action in a way that encompasses group dynamics. For this, we need a new tool. This tool, should be able to sift through the gargantuan amount of data that is collected about human movement, and structure it in a way that we can refer to individual movements, but also to actions, and sub-actions, and sequences of actions, and interactions and plans. This tool will create representations of action at different levels of abstraction. This tool is a Human Activity Language, of the kind envisioned in this paper.

6. Ambient intelligence in the service of artificial intelligence and cognitive systems

Human-machine communication requires partial conceptual alignment for effective sharing of meaning. Concepts are the elementary units of reason and linguistic meaning and represent the cognitive structures underlying phonemes/words. A commonly held philosophical position is that all concepts are symbolic and abstract and therefore should be implemented outside the sensorimotor system**.** This way, meaning for a concept amounts to the content of a symbolic expression, a definition of the concept in a logical calculus. This is the viewpoint that elevated AI to the mature scientific and engineering discipline it is today. Despite the progress, there is still inability of text-based (NLP) technologies to offer viable models of semantics for human computer interaction. For example, imagine a situation where a human user is interacting with a robot around a table of different colored/shaped objects. If the human were to issue the command "give me the red one," or "give me the long one" both the manually-coded and statistical models of meaning employed in text-based NLP are inadequate; for, in both models, the meaning of a word is based only on its relations to other words.

There is however another viewpoint regarding the structure of concepts which states that concepts are grounded in sensorimotor representations. This sensorimotor intelligence considers sensors and motors in the shaping of the cognitive hidden mechanisms and knowledge incorporation. There exists a variety of studies in many disciplines (neurophysiology, psychophysics, cognitive linguistics) suggesting that indeed the human sensory-motor system is deeply involved in concept representations. The functionality of Broca's region in the brain and the mirror neurons theory suggests that perception and action share the same symbolic structure that provides common ground for sensory-motor tasks (e.g. recognition and motor planning) and higher-level activities.

Perhaps the strongest support for the sensorimotor theory comes from the work of Rosch and colleagues [9,25]. The classic view assumed that categories formed a hierarchy and that there was nothing special about the categories in the middle. Take hierarchies like vehicle/car/sports car or furniture/chair/rocking chair [25]. The categories in the middle, according to Rosch, are special – they are the basic level categories. One can get a mental image of a car or a chair but not of a piece of furniture or a vehicle in general. We have motor programs for interacting with chairs, but not with pieces of furniture. In addition, words for basic level categories tend to be learned earlier, to be shorter, more frequent, be remembered more easily, and so on. Thus, the basic level category is the level at which we interact optimally in the world with our bodies. The consequence is that categorization is embodied, given by our interactions.

6.1. Language grounding

In the example given before, in order for the robot to successfully "give me the red one," it must be able to link the meaning of the words in the utterance to its perception of the environment. Thus, recent work on grounding meaning has focused on how words and utterances map onto physical descriptions of the environment: either in the form of perceptual representations or control schemas [1,5,6,26]. Here is the critical point: if we can make a language out of the sensorimotor representations that arise from our actions (in general interactions with our environment), then we can obtain abstract descriptions of human activity that have been obtained from non text (language) data (sensory and motor). These representations are immediately useful since they can ground basic verbs (walk, turn, sit, kick, and so on). It is intuitively clear that we, humans, understand a sentence like "Joe ran to the store" not by checking "ran" with the lexicon (dictionary) but because we have a sensorimotor experience of running. We know what it means to "run", we can "run" if we wish, we can think of "running". We have functional representations of running that our language of action provides.

While such physical descriptions are useful representations for some classes of words (e.g., colors, shapes, physical movements), they may not be sufficient for more abstract language, such as that which denotes intentional action. This insufficiency stems from the fact that intentional actions (i.e. actions performed with the purpose of achieving a goal) are highly ambiguous when described only in terms of their physically observable characteristics. For example, imagine a situation in which one person moves a cup towards another person and says the unknown word "trackot." Now, based only on the physical description of this action, one might come to think of "trackot" as meaning anything from "give cup", to "offer drink", to "ask for change." This ambiguity stems from the lack of contextual information that strictly perceptual descriptions of action provide.

A language of action provides a methodology for grounding the meaning of actions, ranging from simple movement to intentional action (e.g., "walk to the store" versus "go to the store", "slide the cup to him" vs. "give him the cup"), by combining the grammatical structure of action (motoric and visual), with the well known grammatical structure of planning or intentional action. Specifically, one can combine the bottom up structure discovered from movement data with the top down structure of annotated intentions. The bottom up process can give us actual hierarchical composition of behavior, the top down process gives us intentionally-laden interpretations of those structures. It is likely that top down annotations will not reach all the way down to visual-motor phonology, but will perhaps be aligned at the level of visuo-motor morphology or even visuo-motor clauses.

7. Summary

Multi-camera laboratories will become the places where Artificial Intelligence could be redefined by studying meaning through the utilization of both sensorimotor representations and symbolic representations, using machine learning techniques on the gargantuan amounts of data collected. This will lead eventually to the creation of the PRAXICON, an extension of the LEXICON that contains sensorimotor abstractions of the items of the LEXICON [27]. The entire enterprise may be seen in the light of the new emerging *Network Science,* the study of human behavior, not in isolation, but in relation to other humans and the environment.

Acknowledgements

The support of NSF, NIH and the European Union (Cognitive Systems – project POETICON) is gratefully acknowledged. Special thanks to Abhijit Ogale who initiated the study of synergies in our laboratory.

References

[1] M.A. Arbib, T. Iberall, D. Lyons, Schemas that integrate vision and touch for hand control, *Vision, brain, and cooperative computation*, MIT Press, Cambridge, MA, 1987.

[2] J.C. Augusto and P. McCullagh, Ambient Intelligence: Concepts and Applications, Invited Paper by the *International Journal on Computer Science and Information Systems*, volume 4, Number 1, pp. 1-28, June 2007.

[3] J.C. Augusto, Ambient Intelligence: The Confluence of Pervasive Computing and Artificial Intelligence. In *Intelligent Computing Everywhere*, Alfons Schuster (Ed.). Pages 213-234. Published by Springer Verlag, 2007.

[4] J.C. Augusto and D. Shapiro (Eds.). Advances in Ambient Intelligence Volume 164*, Frontiers in Artificial Intelligence and Applications* (FAIA) series, IOS Press. 2007.

[5] D. Bailey, N. Chang, J. Feldman, and S. Narayanan (1997) Extending Embodied lexical development. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society.*

[6] D. Bailey, J. Feldman, S. Narayanan, G. Lakoff, Embodied lexical development, in: Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society, Erlbaum, Mahwah, NJ, 1997.

[7] Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C.A., 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.

[8] Chih-Chieh Han, R. Kumar, R. Shea, E. Kohler, and M. Srivastava. A dynamic operating system for sensor nodes. In *Proceedings of the Third International Conference on Mobile Systems, Applications, And Services (Mobisys),* 2005.

[9] V. Gallese and G. Lakoff, The Brain's concepts: The role of the sensory motor system in conceptual knowledge, *Cognitive Neuropsychology*, 2005, 21.

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

[11] G. Guerra and Y. Aloimonos, Human Activity Language, *IEEE Computer*, May 2007.

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

[13] Guerra-Filho G., Fermuller C., and Aloimonos Y. (2005) Discovering a language for human activity, in *Proceedings of AAAI 2005 Fall Symposium:* "*From Reactive to Anticipatory Cognitive Embodied Systems*".

[14] R. Jackendorf, *The Architecture of the language faculty,* MIT Press, 2000.

[15] O. Jenkins and M. Matarić, "Automated derivation of behavior vocabularies for autonomous humanoid motion", in *Proc. of the International Conference on Autonomous Agents*, pp. 225-232, 2003.

[16] Jeannerod M.: Object oriented action. In *Insights into the reach to grasp movement.* Edited by Bennett K.M.B., Castiello U. Elsevier and North-Holland; 1994:3-15.

[17] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational Markov models. In *Nineteenth International Joint Conference on Artificial Intelligence*, 2005.

[18] D. Lymberopoulos and T. Teixeira and A. Savvides, Detecting Patterns for Assisted Living Using Sensor Networks, *Proceedings of IEEE SensorComm* 2007, October 14-20, Valencia, Spain.

[19] D. Lymberopoulos, A. Ogale, A. Savvides, and Y. Aloimonos. A sensory grammar for inferring behaviors in sensor networks. in the *Proceedings of Information Processing in Sensor Networks, IPSN* 2006, April 2005.

[20] L. Nachman. Intel Corporation Research Santa Clara. CA. New tinyos platforms panel:iMote2. In The Second International TinyOS Technology Exchange, Feb 2005.

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

[22] A. Mussa-Ivaldi, E. Bizzi: Motor learning through the combination of primitives*. Philos Trans Roy Soc Lon Ser B-Biol Sci* 2000, 355:1755-1769.

[23] S. Ogale, A. Karapurkar, and Y. Aloimonos. View invariant modeling and recognition of human actions using grammars. Workshop on Dynamical Vision at ICCV'05, October 2005.

[24] Rao, A. Shon and A. Meltzoff, A Bayesian model of imitation in infants and robots, *Imitation and Social Learning in Robots, Humans, and Animals*, Cambridge University Press, 2005.

[25] Rosch, E., Categorization, in V.S. Ramachandran (Ed.), *The encyclopedia of human behavior*, Academic Press.

[26] D. Roy (2005) Semiotic schemas: a framework for grounding language in action and perception*, Artificial Intelligence,* 167, 2005, 170-205.

[27] The POETICON: The Poetics of everyday life: Grounding resources and mechanisms for artificial agents, Project funded under the 7th Framework, European Union.

[28] Schaal S., Ijspeert A., Billard A.: Computational approaches to motor learning by imitation. *Philos Trans R Soc Lond B Biol Sci* 2003, 358:537-547.

[29] Siskind J. (2001) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15, pp. 31—90.

[30] P. Viviani: Do units of motor action really exist? In *Generation and Modulation of Action Patterns*. Edited by Heuer H., F.C. Springer; 1986:201-216.

[31] T. Teixeira and A. Savvides Lightweight People Counting and Localizing in Indoor Spaces using Camera Sensor Nodes, *Proceedings of the First ACM/IEEE Conference on Distributed Smart Cameras, ICDSC 2007*, October 24-28, Vienna, Austria.

[32] Del Vecchio D., Murray R.M., Perona P.: Decomposition of human motion into dynamics-based primitives with application to drawing tasks. *Automatica* 2003, 39:2085-2098.