Classifying Human Activities in Household Environments

Stefan Vacek, Steffen Knoop, Rüdiger Dillmann
Institut for Computer Design and Fault Tolerance
University of Karlsruhe
D-76128 Karlsruhe, Germany
{vacek, knoop, dillmann@ira.uka.de}

Abstract

The recognition of daily human actvitities is becoming more and more important for humanoid robots. For the robot, being a daily companion of the human, it is crucial that it is able to understand what the human is doing in order to react accordingly.

Although there exist many systems for recognising human activities there is a lack of having a structured classification of these activities. In this paper different concepts for classifying human activities are presented. First, a concept motivated by recognition approaches is presented, which is called structural classification. The second concept is guided by the functional meaning of activities. These two concepts are then combined in a third classification which connects the structural with the functional classification. The benefit of the combined classification is, that it adds semantics into the structural view of activities and it enables the algorithms for further refinement of the recognition.

1 Introduction

One investigated field of application of mobile robots is within the environment of humans. For the robot, being a companion of the human in household environments, it must have several abilities. These include supporting or helping the human, interacting with the human, learning skills and tasks or recognising the human's intention. While the detection and tracking of people is the first step for the robot of being aware of humans in its surrounding, it is important to understand what the human is doing.

The aim of this paper is to investigate concepts for designing a classification of human activities in household environments. This classification supports several research activities and has multiple benefits, which are:

- It serves as a basis for the recognition of activities and can be used several algorithms. Furthermore with this classification it is possible to introduce semantic knowledge into the recognition.
- It establishes a system wide common taxonomy about human activities which can be used widely. Especially for a humanoid robot, this knowledge can be used in:
  - Dialogues, by helping to understand the users actions and recognising his or her willingness to interact.
  - Learning skills and tasks from a human while observing the demonstrator.
  - Situation awareness and intentionality, by understanding the human's activity which is part of the actual situation and recognising his or her intention.
- It helps to build a semantic link between the robot's own abilities and the activities of humans. Thereby it supports the robot to reason about its own abilities and to decide whether and how it can help the human.

It is obvious, that not all possible human activities can be classified. The reasons are that the kind of classification always depends on its purpose and each field of application has its own interests which cannot be covered completely in an overall classification. Therefore, the presented classifications concentrate on a subset of possible human activities: activities in household environments. Its purpose is towards the use in a humanoid mobile robot being a daily companion of the human.

In the following section a brief overview of the state of the art is presented, in section 3 a general introduction on human activities is given and a categorisation of classifying activities is described. The succeeding subsections explain the different concepts of classifying human activities. The combined structure, depicted in section 3.3, incorporates the presented approaches into one classification.

2 Related work

Most of the researchers do not define an explicit classification of human activities. In fact most publications concentrate on detection, recognition and interpretation.

Sierhuis et al. [Sierhuis *et al.*, 2000] describe a representation of work practice which consists of activities of the involved people. Work is defined as transforming input to output. An activity is more than that, namely it includes also collaboration between individuals. An activity is described by how, when, where and why an activity is performed and identify the affects of an activity. Activities locate behaviour of people and their tools in time and space.

In [Rao and Shah, 2001] a flat list of captured actions is used. The recognition evaluates the position of the hand in order to interpret the resulting trajectories. [Sukthankar and Sycara, 2005] uses an acyclic graph to model a specific behaviour. Each edge consists of a basic body motion together with an environmental feature.

Lokman and Kaneko [Lokman and Kaneko, 2004] presented a hierachical structure of the body-parts and joints to derive a classification of human actions. The basic ideas are, that the human does not always use all body-parts for an activity and that multiple actions could happen simultaneously.

A hierarchical structure of actions is used in [Mori *et al.*, 2004] where the actions are classified in a tree-like structure. An action is modelled by Continous Hidden Markov Models. The recognition starts at the root node and for all child nodes, the likelihood is calculated. If there is a valid child, the recognition descends in this lower level and the recognition starts again. If no valid child can be found, the recognition stops. At each level of the tree, there is a special node, called ”etc” which denotes ”every other” action, not listed in the tree at that level. For example at the first level, there are ”Sitting”, ”Lying”, ”Standing” and ”Etc”.

In [Kojima *et al.*, 2002] a concept hierarchy of body actions is used for extracting a natural language description of human actions out of image sequences. An activity is represented by a so called ”case frame” where a case frame expresses the relationship between cases in a natural sentence (like *agent, object, locus, source*, etc.). The hierarchy of actions starts at a generic level and is refined at each level by introducing additional values into the case frame. These additional values correspond to extracted image features. E. g. *be* becomes *move* by introducing the speed of the torso and therefore replacing the verb.

A similar approach is used in [Herzog and Rohr, 1995]. Here, an activity is represented in terms of predicate logic. Each term then describes an action with specific attributes which can be further refined (e.g. ”move” + ”fast” becomes ”running”).

Patterson et al. [Patterson *et al.*, 2003a] use RFID-tags to observe the user’s interaction with objects. The activity models should be human understandable and that they describe the activities intuitively. An activity is described by a set of touched objects. For recognising an activity they use Dynamic Bayesian Networks.

Different aspects of modelling and recognising human behaviours are present in [Liao *et al.*, 2004]. Modelling human behaviour comprises the decomposition of behaviours and the abstraction and thus the grouping of behaviours. A big problem in human behaviour recognition is the gap between the raw sensor data and the recognition algorithms. In [Patterson *et al.*, 2003b] GPS data is used to infer about the user’s movement within a city and his transportation mode (i.e. by bus, by car or by foot). A particle filter is used to estimate the state of the user.

In [Bui, 2003] a framework for probabilistic plan recognition of hierarchies of activities is presented. So called Abstract Hidden Markov Memory Models are introduced which allow to estimate sub-policies depending on the previous history of the process. The system is demonstrated in an office monitoring scenario where different actions like ”going to the printer” are recognised.

Another approach for hierarchical modelling is presented in [Pynadath and Wellman, 2000]. A PSDG (probabilistic state-dependent grammar) is used to define plans and subplans. Parsing a given observation results in probabilities for different subplans allowing the recognition of actions.

3 Classification of human activities

Before defining a classification of human activities, it has to be made clear what the term ”activity” stands for. Following dictionaries (e.g. [dictionary.com, 2005]), they state:

Definition 1 activity*: ”state of being active”*

Looking into the more specific term *human activity*, dictionaries (s. e.g. [WordNet 2.0, 2005]) define it as:

Definition 2 human activity*: ”something that people do or cause to happen”*

It is clear that it is not possible to classify *all* existing human activities. In fact a classification for only a subset, namely activities in household environments, is presented. Beside looking into typical household scenarios, the demands arising from the *COGNIRON* project (the *cognitive robot companion*, s. `http://www.cogniron.com`) were taken into account.

Typical activities are:

- Talking to someone
- Walking around
- Sitting on a chair
- Taking out a beer from the fridge
- Opening a door
- Grasping a cup
- Placing a cup on a saucer

This list isn’t complete, it should only give an impression about the variety of human activities in typical household scenarios. Indeed, these activities can also be combined like *walking while talking to someone*.

For designing the classification, some important issues have to be considered:

- The classification should not depend on any existing algorithm doing activity recognition but it must also be possible to use this classification for the development of future recognition algorithms.
- It should be open ended in a way that new categories could be added in the future and also previously unconsidered activities should be categorised later on.
- It should have a clear structure for the ease of usage.
- It should be usable for different disciplines, like computer vision, dialogues or task learning.

Therefore different concepts of classifications were investigated following different approaches. The first one is derived from the *structure* of the human body, that is, each activity is classified based on the body parts which are *used* for this activity (”How is the activity performed”). The second one is guided by the *functional* meaning of the activities. That is, the semantics of an activity is classified according to its function (”What is the aim of the activity”). Finally these two concepts are incorporated into one where the two previous concepts (structural vs. functional) are orthogonal.

The following subsections describe these concepts in detail.

3.1 Classification by structure

As has been mentioned before, classifying activities by structure means that each activity is categorised based on the involved *structures*. The term *structure* is meant to be a body part, the whole person, an object or a place. The classification is an algorithmic guided approach, because many algorithms evaluate the pose and motion of certain body-parts in order to recognise the activity (e.g. [Aggarwal and Cai, 1999]. It starts with groups of activities belonging to single body-parts and creates new groups by combining groups.

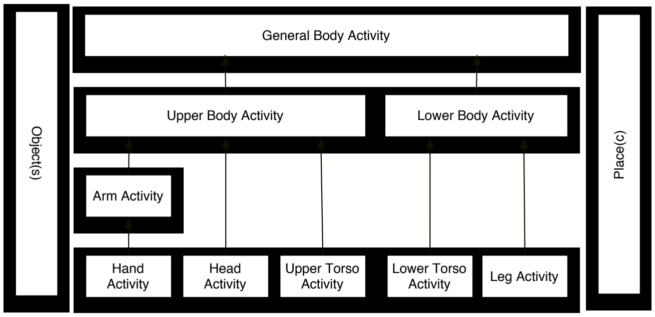

Figure 1: Overview of human activities classified by structure

Figure 1 shows the identified groups of the classification. Each part describes, which structure is involved in this particular group. The arrows, going from one part to another (e.g. from ”Head activity” to ”Upper Body Activity”), denote dependencies of body parts required for this (higher level) group of activities. In the case of ”upper body activity” not all incoming parts need to be active in order to form a valid activity.

The two groups *object* and *place* play a special role in a way that they can augment the meaning of each part. For example, ”hand activity” together with ”object” form a new group of activities. Another example is the command ”put that there” where an object and a place are involved.

It is clear, that a group like ”arm activity” contains a lot of activities like all arm gestures, grasps, etc. Therefore, this classification is a very coarse one, but it helps in understanding how a specific activity is performed or to be more precise, which parts are involved in an activity.

3.2 Classification by function

In contrast to the previously described structural classification, the *classification by function* is guided by the purpose or aim of an activity. In cognitive psychology, human activity is characterized by three features [Anderson, 1989]:

Direction: Human activity is purposeful and directed to a specific goal situation.

Decomposition: The goal that is to be reached is being decomposed into subgoals.

Operator selection: There are known operators that may be applied in order to reach a subgoal. The concept *operator* designates an action that directly realizes such a subgoal. The solution of the overall problem is representable as a sequence of such operators.

Humans tend to perceive activity as a clearly separated sequence of elementary actions. Therefore the set of supported elementary actions is derived from human activity mechanisms. Based on the purpose that is being aimed at by the activity, a classification into two categories is appropriate:

Performative activities: These activities aim at reaching a certain goal in terms of fulfilling a task, they change the state of the human or the state of his or her environment like walking around or grasping an object.

Interaction activities: This class does not only comprise activities within a dialogue, but also for enhancing the learning of demonstrated tasks and guiding the robot.

Figure 2 shows the overall classification based on the modality of their application. Performative and interactive activities are explained in more detail in the following subsections.

Performative Activities

Manipulation, navigation and the utterance of verbal performative sentences are classified as performative activities.

Manipulation: During object manipulation, grasps and movements are relevant for interpretation.

Grasps: For the classification of grasps that involve one hand established schemes can be reverted to. Here, an underlying distinction is made between grasps that do not need to change finger configurations while holding an object until placing it somewhere (”static grasps”) and grasps that require such configuration changes (”dynamic grasps”). While for static grasps exist exhaustive taxonomies based on finger configurations and the geometrical structure of the carried object, dynamic grasps may be categorized by movements of manipulated objects around the local hand coordinate system. Grasps being performed by two hands have to take into account synchronicity and parallelism in addition to single grasp recognition.

Movement: Here, the transport of extremities and of objects has to be discerned. The first may be further partitioned into movements that require a specific goal pose and into movements where position changes underly certain conditions (e.g. force/torque, visibility or collision). On the other hand, the transfer of objects can be carried out with or without contact. It is very useful to check if the object in the hand has or has not tool quality. The

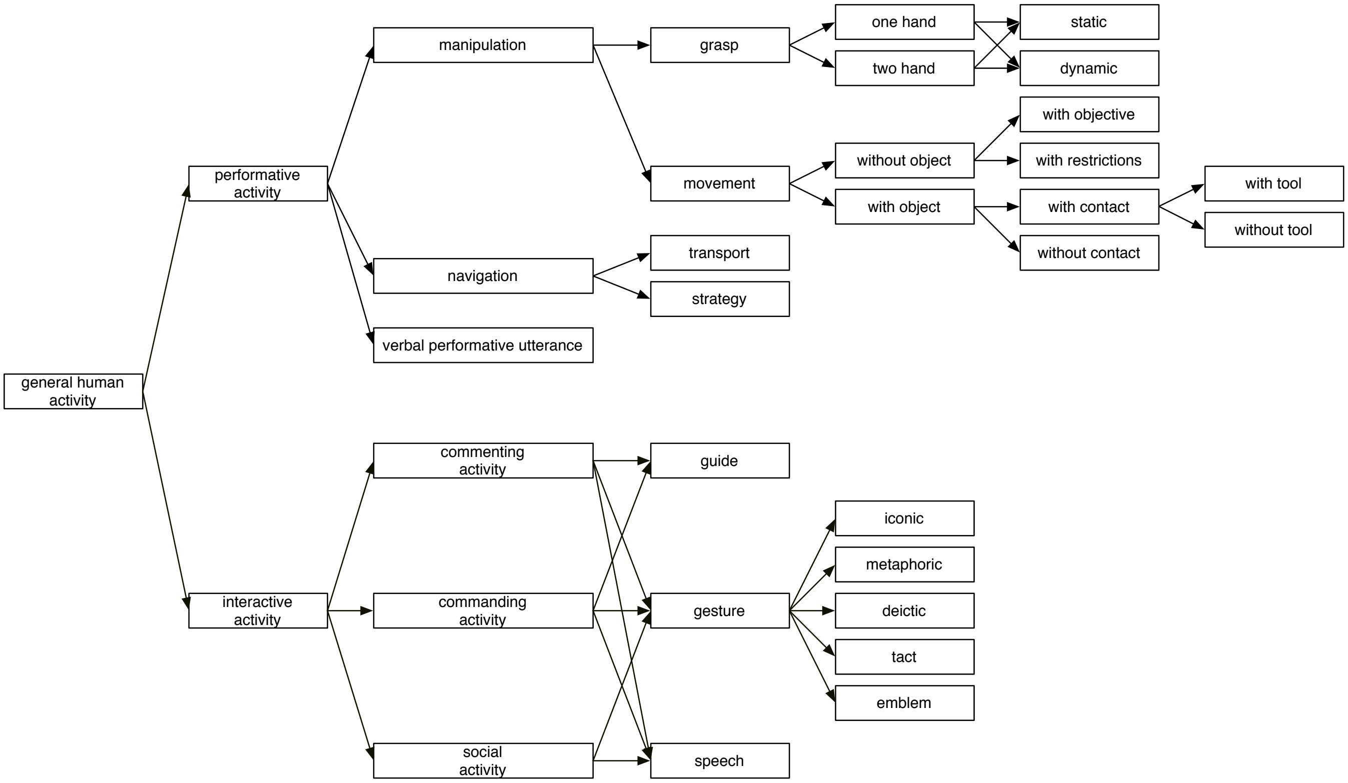

Figure 2: Overview of human activities classified by function

latter case eases reasoning on the goal of the operator (e.g.: *tool type* screwdriver → *operator* turn screw upwards or downwards).

Navigation: In contrast to object manipulation, navigation means the movement of the human himself. This includes position changes with a certain destination in order to transport objects and movement strategies that may serve for exploration.

Verbal performative utterance: In language theory, utterances are performative if the speaker is performing the activity he is currently describing. This could help robotic systems to understand the actual activity.

As can be seen by the complexity of grasp performance or navigation, observation of performative actions requires vast and dedicated sensors. Hereby, diverse information is vital for the analysis of an applied operator: a grasp type may have various rotation axes, a certain flow of force/torque exerted on the held object, special grasp points where the object is touched etc.

Interaction Activities

Commenting, commanding and social interaction are classified as interaction activities. They are not only performed using speech but also gestures with head and hands belong to these categories.

Commenting activities: Humans refer to objects, places and processes by their name, they label and qualify them. Primarily, this type of action serves for enhancing dialogues and it also helps for learning and interpreting.

Commanding activities: Giving orders falls into the second category. This could be e.g. commands to move, stop, hand over or even complex sequences of single commands, that directly address robot or human activity.

Social activities: This class is mainly intended at exchanging information. It includes activities like greeting or asking.

It is clear, that in contrast to the structural classification, a single activity of a body-part can result in activities of different groups. For example an activity with the hand can be a grasping activity or a commanding gesture. Furthermore, a single activity can be achieved with different body-parts, for example affirmation (a commenting activity) can be done with the hand (”thumbs up”) as well as with the head (”nodding”).

3.3 Combining the classifications

The problem of the two presented classifications is, that they consider mainly one dimension of concepts for classifying human activities. To be more precise, the structural classification is based on *how* (or *which body-part*) an activity can be detected and therefore classified. On the other hand, the functional classification mainly concentrates on *what* type of activity is present without considering which body-parts are involved (and therefore need to be algorithmically evaluated).

So the question is how to create a classification which connects the *how* and the *what* type. Or, in other words, how to fill the gap between semantic and detection.

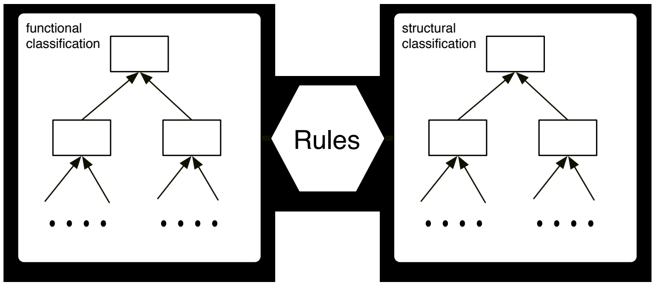

Figure 3: Connecting the structural with the functional classification.

The idea is to introduce a set of rules (s. figure 3) which connects certain structural parts with the corresponding functional group. The connection is bi-directional giving information of the structural into the functional classification and vice versa.

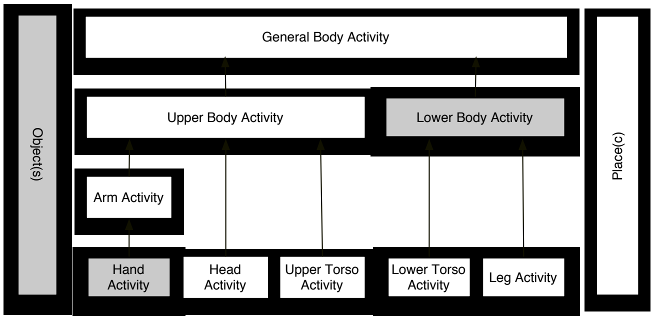

Figure 4: The relevant structural groups for the example ”transport of an object”.

We illustrate the idea using the activity ”transporting an object” as an example. In the structural classification, the groups ”Object(s)”, ”Hand Activity” and ”Lower Body Activity” are involved, which can be detected with appropriate sensors. This is depicted in figure 4. The set of rules, which holds the background knowledge about mapping between structural and functional representations, activates the corresponding activities in the functional model.

At this point, the hierarchical form of the functional classification enables further reasoning about the performed activities and their more generalised activity classes. In the given example, it is now possible to derive the classes ”one hand”, ”grasp”, ”manipulation”, etc. The advantage of having the more generalised classes is, that others could use the information at the level of detail they need. For example, if the dialogue only wants to know, if there is a performative activity, the requested information can easily be delivered.

Additionally, the knowledge of the functional classification allows also for refining the detected activity. More features can be extracted by the perception in order to evaluate if there is a more specific activity. Also, the current context can be used for further refinement.

4 Conclusion and Future Work

In this paper we presented different concepts for classifying human activities. The idea is to establish a common taxonomy for recognising activities as well as using it in other applications. The classification was done for activities in household environments to help humanoid robots in recognising human activities. The first classification is based on the body structure of the human being, which is also motivated by algorithmic approaches. The second classification is structured based on the functional meaning of a human activity, bringing semantics into the classification. These two classification are then combined into a third classification which connects the structural (body-part driven) view with the functional view.

The next steps are to further validate the proposed classification and to continue with the classification in terms of extending it with new activities. For validating the classification an activity recognition will be developed which will be used to teach the robot and to detect the users intention in order to enable the robot to assist the human. Additionally, social studies about human behaviour in the presence of robots will be investigated. The set of rules will be developed in order to establish the connection between the structural and the functional classification. Furthermore, investigations will be done, how the rules can be learned in order to reduce the required a priori knowledge.

Acknowledgments

The work described in this paper was conducted within the EU Integrated Project COGNIRON (”The Cognitive Robot Companion”, s. `http://www.cogniron.com`) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

References

[Aggarwal and Cai, 1999] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999.

[Anderson, 1989] J. Anderson. *Kognitive Psychologie, 2. Auflage*. Spektrum der Wissenschaft Verlagsgesellschaft mbH, Heidelberg, 1989.

[Bui, 2003] Hung H. Bui. A general model for online probabilistic plan recognition. In *International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, August 9-15 2003.

[dictionary.com, 2005] dictionary.com. Definition of activity, 2005.

[Herzog and Rohr, 1995] Gerd Herzog and Karl Rohr. Integrating vision and language: Towards automatic description of human movements. In *Proc. of the 19th Annual German Conference on Artificial Intelligence (KI-95)*, Bielefeld, Germany, Sept. 11-13 1995.

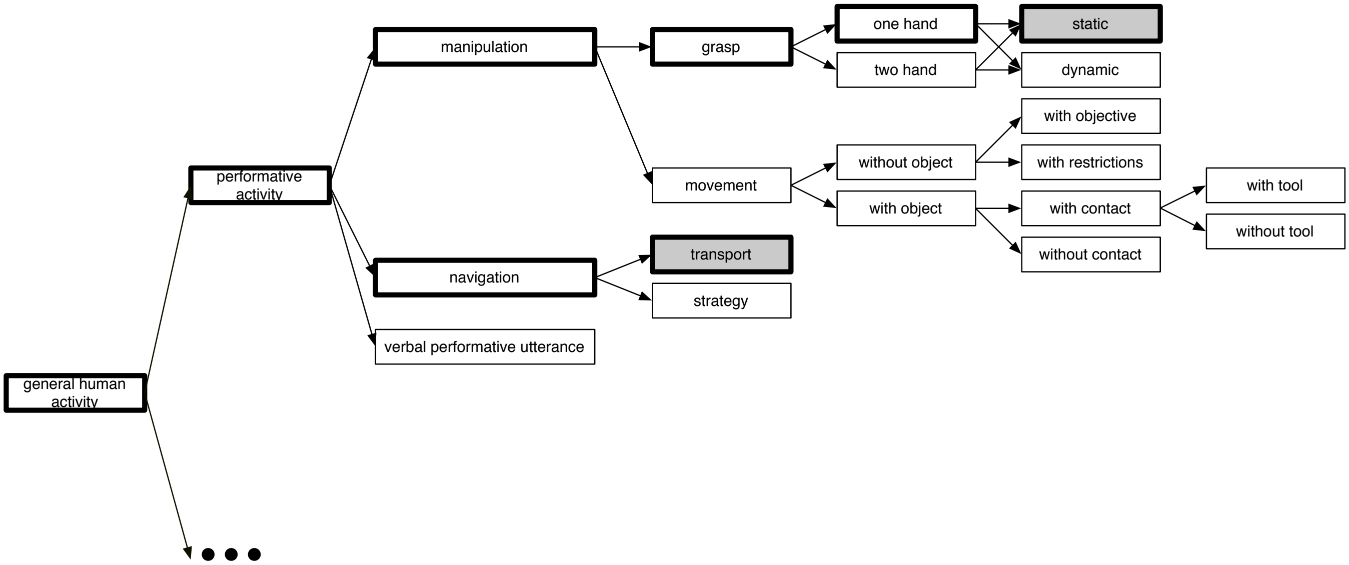

Figure 5: The relevant structural groups for the example ”transport of an object”.

[Kojima *et al.*, 2002] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activites from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 2(50):171–184, 2002.

[Liao *et al.*, 2004] Lin Liao, Don Patterson, Dieter Fox, and Henry Kautz. Behavior recognition in assisted cognition. In *The AAAI-04 Workshop on The AAAI-04 Workshop on Supervisory Control of Learning and Adaptive Systems*, San Jose, California, USA, July 25 2004.

[Lokman and Kaneko, 2004] Juanda Lokman and Masahide Kaneko. Hierarchical interpretation of composite human motion using constraints on angular pose of each body part. In *13th IEEE Int'l Workshop on Robot and Human Interactive Communication (RO-MAN 2004)*, pages 335–340, Kurashiki, Okayama Japan, Sept. 20-22 2004.

[Mori *et al.*, 2004] Taketoshi Mori, Yushi Segawa, Masamichi Shimosaka, and Tomomasa Sato. Hierarchical recognition of daily human actions based on continous hidden markov models. In *Proc. of the Sixth IEEE Int'l Conf. on Automatic Face and Gesture Recognition (FGR'04)*, Seoul, Korea, May 17-19 2004.

[Patterson *et al.*, 2003a] Don Patterson, Dieter Fox, Henry Kautz, and Matthai Philipose. Expressive, tractable and scalable techniques for modeling activities of daily living. In *Proceedings of UbiHealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications*, Seattle, Washington, USA, October 12, 2003 2003.

[Patterson *et al.*, 2003b] Donald J. Patterson, Lin Liao, Dieter Fox, and Henry Kautz. Inferring high-level behavior from low-level sensors. In *Proc. of the International Conference on Ubiquitous Computing (UbiComp)*, pages 73–89, Seattle, Washington, USA, October 12-15 2003.

[Pynadath and Wellman, 2000] David V. Pynadath and Michael P. Wellman. Probabilistic state-dependent grammars for plan recognition. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, pages 507–514, Stanford, California, USA, June 30 - July 3 2000.

[Rao and Shah, 2001] Cen Rao and Mubarak Shah. View-invariance in action recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, volume II, pages (II)316 – (II)322, Kauai Marriott, Hawaii, Dec. 9-14 2001.

[Sierhuis *et al.*, 2000] Maarten Sierhuis, William J.Clancey, Ron van Hoof, and Robert de Hoog. *Modelling and Simulating Human Activity*. AAAI Fall Symposium on Simulating Human Agents., 2000.

[Sukthankar and Sycara, 2005] Gita Sukthankar and Katia Sycara. A cost minimization approach to human behavior recognition. In *Proceedings of Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, to appear 2005.

[WordNet 2.0, 2005] WordNet 2.0. Definition of human activity (http://www.cogsci.princeton.edu/cgi-bin/webwn2.0?stage=1&word=human+activity), last visited: 2005/01/11, 2005.