FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D2.31
Report on a classification of human activities in households environments

Due date of deliverable: 31/12/2004
Actual submission date: 31/01/2005

Start date of project: January 1st, 2004 **Duration** : 48 months

Organisation name of lead contractor for this deliverable:
University of Karlsruhe

Revision: final

Executive Summary

This report presents result on classifications of human activities in daily household environments. The classification has several purposes serving as a common taxonomy for all research areas within the project. Additionally it builds a starting point for the recognition of activities on the one hand and for learning skills and tasks on the other hand.
The report starts with on introductory overview why such a classification is important and how a classification can be derived. Two important aspects are taken into account, which influence the development of classification. These are, an algorithmic point of view, having in mind *how* an activity could be recognised and the semantic point of view asking *what* is the common meaning of a certain class of activities. It is clear, that not all existing activities can be classified., therefore the main focus is laid on human activities in household environments.
The first concept of classification is called a classification by structure. It is based on the structure of the human body and the possibly involved structures ”object” and ”place” which define a certain class of activities.
The second concept is called classification by functionality. In contradiction to the first one, it groups activities based on their functional meaning (their semantics). Two main groups are identified: performative activities and interactive activities which are then refined.
Finally, these two approaches are combined, because there has to be a connection between the structural view of activities and the semantic view.

Role of Classification of human activities in Cogniron

A robot which serves as a Cognitive Robot Companion must be able to on the one hand safely navigate and act in the presence of humans, and on the other hand to interact with humans in a natural way. Also, the robot should not only be able to obey and execute direct commands, but also to offer help when needed. Therefore, it must be able to understand human behaviours and activities, which enables the robot to predict actions, intention and goal of the human.
Recognition of human activities is needed as input functionality for several research areas within the project: Before engaging in a dialog (RA1) with the human, the robot has to decide whether it is appropriate to disturb him. During the dialog, it is useful to observe the human’s actions and to use his actions as additional input for interpretation.
While the robot observes the human performing a task in order to learn from it (RA4), it is also crucial to extract as much knowledge as possible. Learning may be enhanced by estimation of the users’s intention, and commenting actions like gestures help by clarifying the current context.
When the robot has to collaborate with a human to solve a problem (RA6), it is most important for the robot to know exactly the goal and intention of the human. To predict this, it is a prerequisite to know about the actions and activities of the human. This is even more crucial when the robot offers his help.

Relation to the Key Experiments

For Key Experiment 1 (”Robot Home Tour”), it is very important that the robot is able to detect the human’s attention. After getting the robot’s attention, the human shows it around the house and points to several objects and spaces. Here, the robot must be able to detect several activities like interaction activities, movements and to dereference objects and spaces which the human indicates (e.g. by

pointing). For KE1 and KE2 (”Curious Robot”) it is essential for the robot to detect and interpret human activites and postures. Especially in KE2, the robot has to engage in a dialog by detecting that the human needs help. This requires knowledge about the human’s intention, which can not be derived without an analysis of the human’s activities.

1 Report on a classification of human activities in household environments

1.1 Introduction and Overview

Being a companion of the human in household environments the robot 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 report 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 which will be developed within the project. 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 COGNIRON, 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 he 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 classification concentrates on a subset of possible human activities in household environments and its purpose is towards the use in a mobile robot.
In the following section a brief overview of the state of the art is presented, in section 1.3 a general introduction on human activities is given and a categorisation of classifying activities is described. The succeeding subsections explain our concept of classifying human activities. The matrix structure, depicted in section 1.6, incorporates the different approaches into one classification.

1.2 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. [9] describe a representation of work practice which consist 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 describe 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 [8] a flat list of captured actions is used. The recognition evaluates the position of the hand in order to interpret the resulting trajectories.
Lokman and Kaneko [6] 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 [7] 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 [5] a concept hierarchy of body actions is used for extracting a natural language description of human actions. 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 [4]. 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”).

1.3 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. [3]), they state:

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

Looking into the more specific term *human activity*, dictionaries (s. e.g. [1]) 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 investigated. Beside looking into typical household scenarios, the demands arising from the scenarios within the project (both from the *cogniron functions* and the *key experiments*) 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 be possible to use this classification for the development of future recognition algorithms.
- It should have a clear structure for the ease of usage.
- 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.

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.

1.4 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. It starts with groups of activities belonging to single body-parts and creates new groups by combining groups. 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.

1.5 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 [2]:

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.

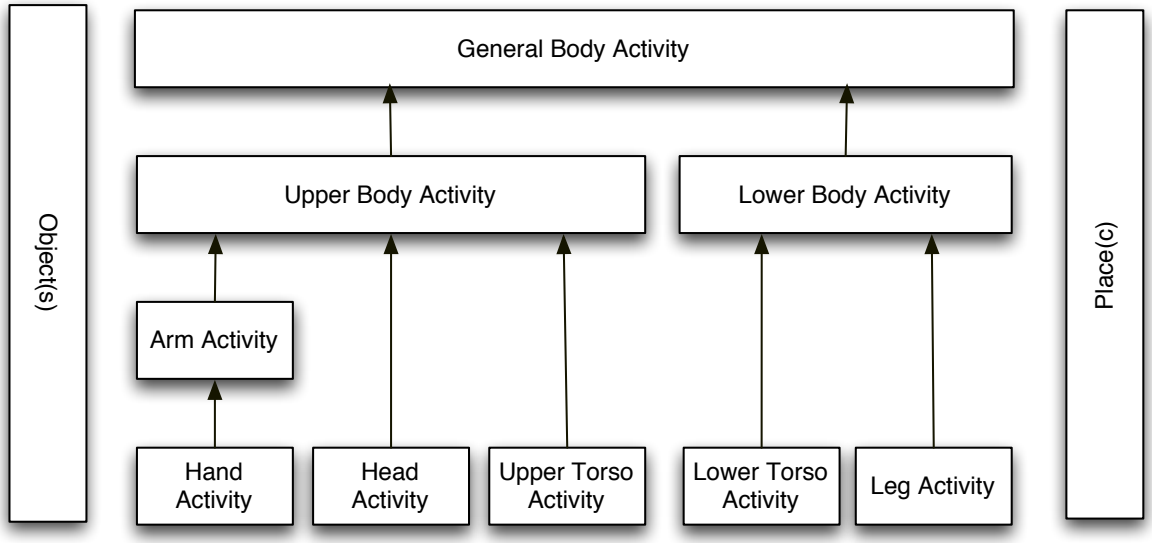

Figure 1: Overview of human activities classified by structure

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.

1.5.1 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

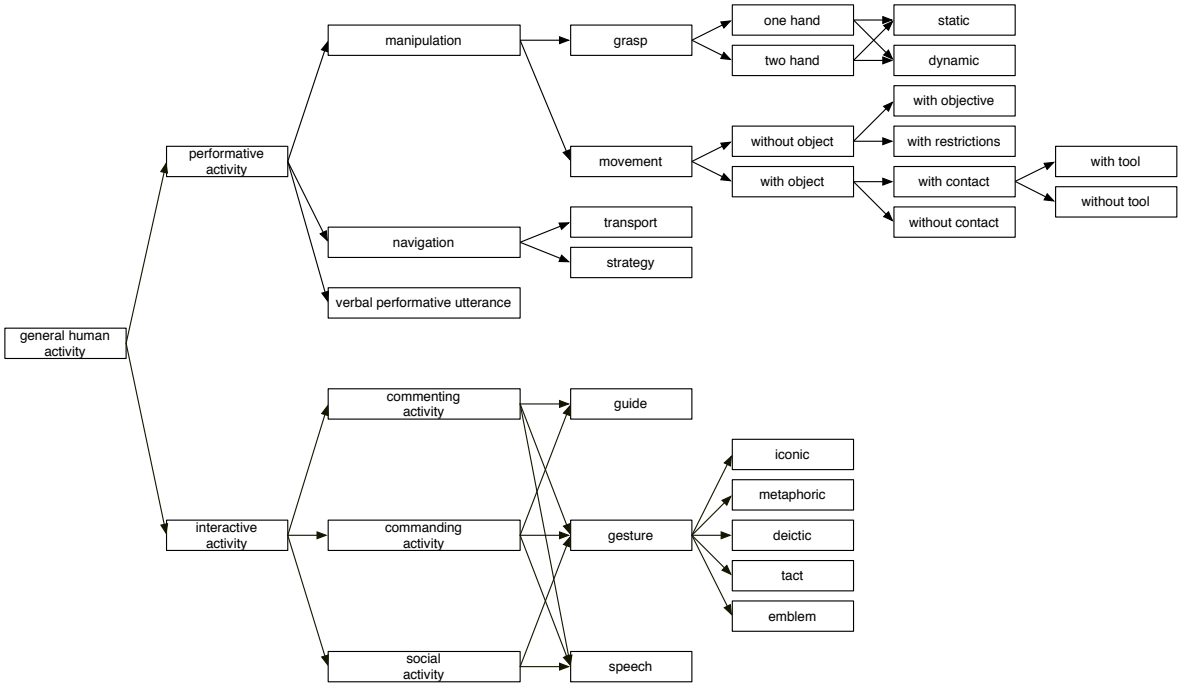

Figure 2: Overview of human activities classified by function

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 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.

1.5.2 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.

1.6 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.

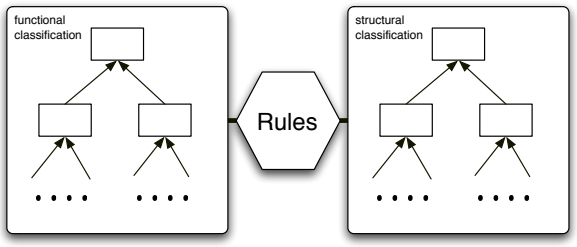

Figure 3: Connecting the structural with the functional classification.

In our approach we introduce a set of rules (s. figure 3) which connects certain structural parts with the corresponding functional group.
We illustrate the approach 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.

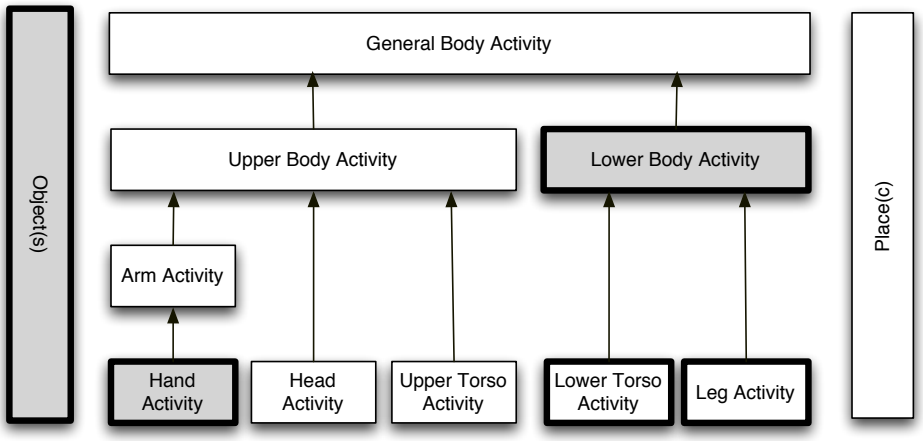

Figure 4: The relevant structural groups for the example ”transport of an object”.

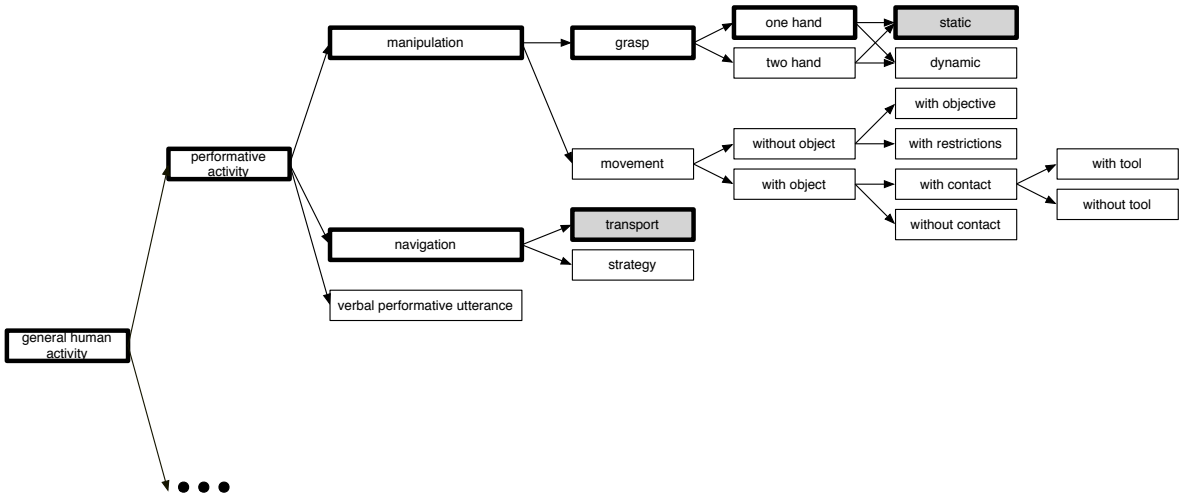

Figure 5: The relevant structural groups for the example ”transport of an object”.

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.

1.7 Applicability

In order to recognise human activities, both human model configurations and movements of human body parts or of the whole human will be considered.
A model configuration is meant to be a static pose of the human which is in geometric terms described by fixed angles of single human body parts, e.g. ”sitting” is described by a specific relation between the upper torso and the legs.
Other activities require a description over time, e.g. ”nodding” is described by a certain angular movement of the head.

The geometric information about the human can be extracted from the human model developed in WP2.2 which includes information about the pose and the movement of each body part. The movements can be extracted from the provided history of the human model.
Based on these poses and movements methods for classification will be used which detect basic activities. The basic activities correspond to the lowest level of the structural classification (s. section 1.4). The structural classification allows to combine these basic activities to extract higher level activities. The combined classification (s. section 1.6) provides a map from the structural to the functional view of human activities.

2 Future Work

The next steps are to further validate the proposed classification and to continue with the classification in terms of extending it with new activities. 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.

3 References

References

[1] 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.

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

[3] dictionary.com. Definition of activity (http://dictionary.reference.com/search?q=activity), last visited: 2005/01/11.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.