Distribution and Recognition of Gestures in Human-Robot Interaction

Nuno Otero, Steffen Knoop, Chrystopher L. Nehaniv, Dag Syrdal, Kerstin Dautenhahn and Rüdiger Dillmann

Abstract— This paper presents an approach for human activity recognition focusing on gestures in a teaching scenario, together with the setup and results of user studies on human gestures exhibited in unconstrained human-robot interaction (HRI). The user studies analyze several aspects: the distribution of gestures, relations, and characteristics of these gestures, and the acceptability of different gesture types in a human-robot teaching scenario. The results are then evaluated with regard to the activity recognition approach.

The main effort is to bridge the gap between human activity recognition methods on the one hand and naturally occuring or at least acceptable gestures for HRI on the other. The goal is two-fold: To provide recognition methods with information and requirements on the characteristics and features of human activities in HRI, and to identify human preferences and requirements for the recognition of gestures in human-robot teaching scenarios.

I. INTRODUCTION

Robots are starting to leave the confines of industrial settings and are moving to highly dynamic and socially challenging human environments. As robots start acting in human social environments issues of agency, believability and sociality become very important [1]. Humans will expect that robots inhabiting their social spaces will conform as much as possible to their expectations. Thus, it seems critical that the interactions need to be "acceptable" and "comfortable" to humans [1]. Fong, Nourbakhsh and Dautenhahn [2] state that the design of sociable robots needs input from research concerning social learning and imitation, gesture and natural language communication, emotion and recognition of interaction patterns.

In relation to human communication, to a large extent, it happens through humans' use of physical movement of their limbs. Many activities can be readily recognized just by observing the motion of the limbs of the human body, or even the motion of the entire body. Additional information for determining a person's current activity can be derived from the environment. This information is often called *context*. Together, body motion and context provide in many situations enough information to derive the person's current activity or activities.

Interpretation of a person's motion within its environment can enhance HRI in different ways: On the one hand, the current action of the human interaction partner can help the robot to plan its own tasks and goals, e.g. in cooperative tasks or for taking decision on becoming proactive. On the other hand, meaningful body motions, that can be designated as *gestures* generally make up a large part of the information flow during interaction. In fact, in humans, gestures are closely linked with the accompanying speech in terms of timing, meaning and communicative function (see, for example, [3][4][5][6][7]). Furthermore, for a robot, it seems plausible to assume that accurate activity and gesture recognition can facilitate the task of its speech recognition system and vice-versa.

This paper outlines our effort, as part of the research for the European funded project COGNIRON, to create a system, able to be incorporated in a robot, for the recognition and interpretation of human activities, including observed body motions and gestures. Two streams of research are being pursued in complementary manner: (a) the technological development of a classification and interpretation system for human gestures and activities, and (b) user studies to capture the corresponding system requirements from the human's point of view. This work was inspired by previous research in COGNIRON, specifically: Nehaniv et al. [8] provided the conceptual framework for the coding scheme categories to classify observed gestures in human activities and research by University of Karlsruhe [9] in relation to requisites for human activity descriptions from a system's perspective.

In terms of the technological part, the paper describes an approach to recognize human activities from the observed motion of the user's body, based on an articulated body model (for details, see [10][11]). From the continuous motion trajectory, features are extracted and evaluated with respect to their relevance for the recognition of a certain activity. Additional features gathered from the environment (*context knowledge*) are included. Activities are then recognized based on the most relevant feature set using a neural network classifier.

In relation to the user studies, the on-going development of a coding scheme to classify gestures people produce when asked to demonstrate how to perform a task will be described. The coding scheme is an essential part of our strategy to systematically study the frequency, duration and sequence of different gestures in people's task demonstrations. Some relevant results concerning the characteristics and distribution of gesture used in human-robot teaching from two user studies will be described.

The paper will also discuss our integration effort concern-

ing the two streams of research and future developments.

II. State of the Art

Good overviews to the area of human activity recognition are the comprehensive surveys composed by Cédras and Shah [12] and by Gavrila [13], and the slightly shorter review by Aggarwal and Cai [14]. Additionally, Wang, Hu and Tan [15] cover some work done after 2000.

A large field of application for activity recognition is given by the problem of video surveillance, e.g. in public areas where often surveillance cameras already exist. The topic of surveillance raises already several restrictions, which include usage of cameras only, or large distance between sensor and target. In most surveillance cases, the background can even be assumed to be static, which is an important factor for recognition and can also affect the method and algorithm selection. Activity recognition systems for surveillance applications have been developed e.g. by Ribeiro et al. [16][17] and Nascimento et al. [18], which rely on large area camera images. These approaches use the trajectory of the whole person for activity classification, and do not use the body configuration. This is on the one hand due to the fact that the observed activties are large-scale and can thus be classified by observing the whole body trajectory over time, on the other hand, the sensor data simply can not provide such detailed information as would be needed for determination of the body configuration. Also, in a surveillance context, the system is only directed at observation, without any active components.

These vision-based activity recognition approaches still follow a common methodology: From the raw input data (camera images in this case), a set of features is extracted which is in a second step processed to classify performed human activities. Although the aim and conditions are different for the context of this work, it still follows a similar approach.

Following this strategy, we divide the process into three steps. The first is *motion capture*, which covers the entire problem of observing a human subject and obtaining a digital representation. The second is *motion analysis*, where the motion capture data is processed to make it suitable for the third step, the actual *recognition* or *classification* itself.

An autonomous robot interacting with humans and moving through a dynamically changing human-inhabited environment presents significant challenges for recognition of human activity and for making use of such recognition to guide its interactive behavior in real time.

III. gesture recognition system framework

In general terms, the proposed activity recognition method consists of the following 3 steps:

1) Observation of the human subject and environment (*motion capture* and other perception modules). This can either be done with view-based approaches, retrieving abstract feature patterns, or based on kinematic models of the human, which results in 2D or 3D motion trajectories of the human body degrees of freedom (DoFs).

2) Analysis of the motion and pattern construction (*motion analysis*). There are two main approaches: Segmentation into basic elements, often called *motion primitives* by recognition of key points, or concurrent feature extraction for a later analysis.

3) Comparison with previously stored patterns (*recognition*). This task is commonly solved using a classifier, which in most cases has been trained with a manually segmented data set. This classifier can either include an explicit time model (like HMMs, Bayesian nets) or be time-independent (like Neural Networks, Support Vector Machines etc.).

For the proposed approach, a *human motion capture* system gathers data of the human configuration over time, resulting in trajectories for each modelled limb and joint angle of the human body in 3d. The motion capture system called *VooDoo* is described in detail in [10], [11]. The used body model and an example configuration are shown in fig. 1.

Fig. 1.   Human body model and motion capture example

¿From this information, for each activity which has to be recognized a set of *model intrinsic features* is derived. These features do not rely on temporal segmentation, but are generated continuously. In addition, a set of *extrinsic features* can be taken into account. These features must be generated by external modules by observation of the environment. The feature synthesis is described in sec. III-A, the evaluation and selection process in sec. III-B.

The *classification step* is performed by a simple Feed Forward Neural Network (FFNN) which processes the feature stream. This FFNN has been trained with manually segmented training examples. It is described in detail in sec. III-C.

The whole process is depicted in fig. 2.

A. Feature extraction

Three types of features are used within this context: Raw model data, filtered model data, and extrinsic features.

1) Raw tracking data: Obviously, the *raw tracking data* consisting of the whole body configuration trajectory can be used as *primitive feature set*, as it contains all available information about the human motion. Raw data can also be combined to retrieve more sensible features: Computation of the Tool Center Point (TCP) height with respect to head height can e.g. help much in separation of different activities.
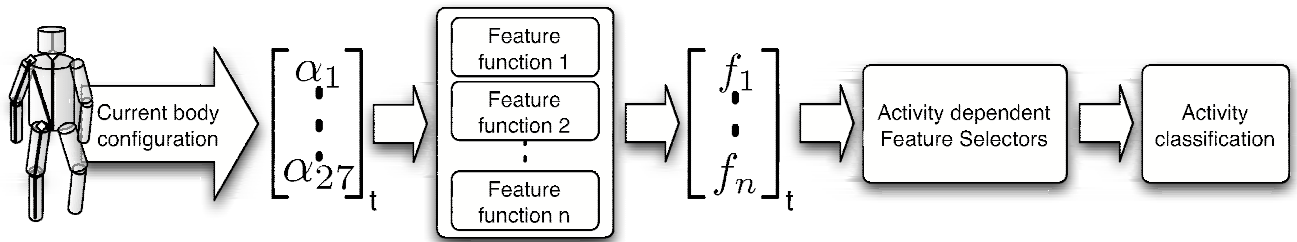
Fig. 2. Feature extraction and activity recognition process. From the body model trajectory, joint angles α_i are extracted for each frame at time t. These angles serve as input for n feature functions, which result in a feature vector $\vec{f_t}$ for each frame. This feature vector is then used by the classifier to recognize activities.

2) Statistical analysis: A high number of statistical methods exist that can be used to extract sensible features from a data set. Our features set contains the following:

- *Covariance* between different input data vectors,
- *Principial Component Analysis* to detect the most relevant features in the current set, or to detect major motion directions (e.g. motion planarity),
- *Frequency analysis* to detect periodic motions.

This reveals a major difference in feature properties: Features can either be *time-dependent* or *snapshots*. Frequency analysis of e.g. elbow motion must be determined using a time window, thus taking motion history into account. TCP height with respect to the head is independent from history. This shows that different classifiers may prefer different feature sets: A classifier which itself takes history into account does not rely on time-dependent features as strong as a classifier without time modeling.

3) Extrinsic Features: Many activities are very simple to recognize when not only the body motion of the human is given, but also information from the environment. This may comprise information about time of day, grasped objects, the current location, or other persons involved.

Generally speaking, this is context information. As the definition of context given e.g. by [19] comprises *all* relevant information from the environment, only part of the context can (and must) be included for activity classification, because (a) only part of the world state can be measured, and (b) only part of all possible human activities is of interest. This included context knowledge is referred to as *extrinsic features*.

In some cases history of recent (or more remote) past interaction may be required to distinguish amongst possible human activities in recognition [8].

B. Feature Evaluation and Selection

Given a large feature set designed to capture a wide variety of activities, we should expect many features to be irrelevant for any given activity. In practice, these features will contribute nothing but noise to the classifier.

Mathematically, the most general statement we can make about a *relevant* feature variable F_i and a target class C is that for a given class value c and at least one feature value f_i, it satisfies the relation

$$p(c = C|f_i = F_i) \neq p(c = C). \quad (1)$$

Two different approaches exist to estimate the relevance measures of a given set of features and classes: Estimation only from sample data, often called *filter methods*, and estimation with the help of an embedded classifier called *wrapper methods* [20].

Filter methods typically make use of various statistical means to evaluate the relevance of features. These are e.g. *Correlation Analysis* and the more general *Mutual Information Analysis*, sometimes also called *Information Gain*.

Mutual information uses the concept of entropy to describe a measurement of the information that is common between two stochastic variables. The mutual information I between the stochastic variables X and Y is defined as:

$$I(X;Y) = \sum_{x,y} P(X = x, Y = y) \log_2 \frac{P(X = x, Y = y)}{P(X = x) \cdot P(Y = y)} \quad (2)$$

Based on the feature evaluation methods, these can now be appropriately selected from the whole set. The goal of feature selection is to find a minimum number of features to achieve the highest possible accuracy. These two goals are commonly mutually exclusive, so this must be weakened to finding a sufficiently small set of features that provide satisfactory prediction.

The problem of finding a minimum set of features is known to be NP-hard in the general case. *Exhaustive search* is a candidate method certain to find the optimum solution. A simplification of this is the *greedy selection*.

Battiti [21] reinterpreted the feature selection problem in terms of information theory for a classification C:

Given an initial set F with n features, find the subset $S \subset F$ with k features that minimizes the conditional entropy $H(C|S)$, i.e. that maximizes the mutual information $I(C;S)$.

Using this interpretation, he proposed a greedy algorithm called *Mutual Information Feature Selector (MIFS)* to select features based on their mutual information with the target class $I(C;F_i)$. The function $I(C;F_i)$ can be estimated by calculation of the entropy and mutual information (MI) using eq. 2.

Kwak and Choi [22] suggested a slightly different measure, and finally give an improved measure $R_2(F_i)$ for the

contribution of a new candidate feature F_i given an already selected set S of features:

$$R(F_i) = I(C; F_i) - \beta \sum_{F_s \in S} \left(\frac{I(F_s; C)}{H(F_s)} I(F_s; F_i) \right), \quad (3)$$

According to [22], the value β gives flexibility to the *MIFS* algorithm. Setting β to zero ignores all mutual information between different input features, and selects features only based on their mutual information with the target class. Setting $\beta = 1$ incorporates the mutual information measure between different input features and thus deselects redundant features.

The proposed approach uses the *MIFS* algorithm for feature selection.

C. Activity Classification

For the classification, a *Feed Forward Neural Network* was chosen. For each activity which has to be recognized a solitary neural network (NN) has been used. So for recognition of n activities, n neural networks exist. Each NN consists of 3 layers with

- k_i input neurons, with k_i the number of selected features for the given activity class c_i. Typical values are $3 \leq k_i \leq 10$.
- 1 output neuron for the current estimation for activity class c_i.
- 10 neurons in the hidden layer. This has been chosen from experiments and experience. The number of selected features gives for all cases $k_i \leq 10$. Less than 10 neurons in the hidden layer decreases recognition results, while choosing more than 10 did not notably increase recognition rates in laboratory tests.

D. Open Questions

To optimally design and parameterize the depicted recognition approach, several issues need to be investigated.

- The main question concerns the set of activities which need to be recognized. Even if the system is able to recognize and classify a large number of activities, reliability and uniqueness increase with decreasing number of different activities. So an optimal set of activities only contains those which are occuring in the given interaction context and which are relevant for the robot for interpretation.
- The human activity model needs to be further refined. This includes not only time-dependency, but also dependencies and transition probabilities (e.g. gesture repetition) between activities and activity classes. Understanding these relations improves not only recognition, but also interpretation of observed activities.
- It is necessary to obtain a clear and representative training set for each activity. The training set has to approximate the whole variety of possible instances likely to occur as well as their characteristics and distribution. Therefore, prior to capturing the training set, one needs a clear understanding of properties of each activity class.

- The time window for recognition of time-dependent features (see sec. III-A.2) must be selected. Provided that good features for a target activity occur during the whole activity (which is especially true for e.g. periodicity, planar motion etc.), the window size depends mainly on the duration of the recognized activities.

So the main questions we impose are *which activities and gestures* have to be recognized, what are the *relations between activities*, what does the *training set* look like and what are the *temporal attributes* of these gestures and activities.

It is obvious that these problems must be solved by studying the behavior of people who actually perform the gestures and activities within a similar context as during recognition.

IV. THE USER STUDY SETUPS

Two exploratory user studies motivated by the above considerations were run to illuminate which naturally occurring gestures can be observed in a scenario specifically relevant to human-robot interaction scenarios for the project (for a summarized review of the domain see [23]). The term 'naturally' here refers to an unconstrained scenario where subjects were not given any scripts or pre-defined gestures to use.

In both studies a within-subjects design was followed and the general task the users had to perform was similar: they had to demonstrate how to lay a table to a video-camera that posed has the vision system of a robot. The second study, however, was not a mere replication of the first - they differed in crucial aspects.

In the first study, two steps were needed to accomplish the main experimental task: first, the 9 participating subjects were asked to gesture for a robot to perform a particular task and second they were requested to actually demonstrate how it should be completed, meaning they would manipulate the objects (for a detailed description of the study see [24]). The task involved: (a) taking some plates from a table, (b) setting the plates and corresponding cutlery in another table and finally (c) pick up the plates again and put them away. The subjects were instructed that only one object could be manipulated at a time. Furthermore, a software program was created to simulate the robot's feedback to subjects. The feedback was simulating the robot's understanding of the gestures produced . The feedback consisted on the display of three colors in a computer screen: a) red if the system did not understand at all the meaning of the gestures produced; b) yellow if the system understood partially but further specification was needed; and c) green if the gestures were understood. However, the actual display of the feedback was random following a random probability distribution of: 20% for red, 20% for yellow and 60% for green. One of the experimenters was controlling the segmentation of when to display feedback by pressing a button at the end of each sequence of the participants' gestures.

In the second study, the 10 participating subjects had to demonstrate how to lay a table for two people utilizing two different methods: using only gestures or gestures and speech

- these were the two experimental conditions (for a detailed description of the study see [23]). Differently from the first study, the subjects did not need to follow the two step rule of miming and demonstrating and no feedback was given.

A. The User Study Coding Schemes for the Classification of Gestures

For the present studies we followed the functional classification system proposed by Nehaniv et al. [8] for the elaboration of a coding scheme to identify people's gestures when asked to explain a home task to a robot. Nehaniv et al. [8] propose the following five functional classes of gestures:[1]

- *Irrelevant and Manipulative* Gestures - these are gestures do not have a primary communicative or interactive function (in practice, this class is split). The former subclass is not relevant for most HRI purposes, but exclusion from consideration following recognition is desirable. To the contrast, manipulative gestures change the environment or human's relation to it.
- *Side Effect of Expressive Behavior* - these are gestures that occur as side-effects of people communicative behavior. It can be motion with hands, arms, face, etc but without specific interactive, communicative, symbolic or referential roles.
- *Symbolic* Gestures - these are gestures that follow a conventionalized signal. Its recognition is highly dependent on the context (both current task and cultural milieu).
- *Interactional* Gestures - this category classifies gestures used to regulate interaction with a partner. Thus they can be used to initiate, maintain, invite, synchronize, organize, regulate, or terminate an interaction behaviour between agents.
- *Referencing/Pointing* gestures (*deixis*) - the gestures that fall into this category are gestures used to indicate objects or loci of interest.

For the first study, the coding scheme used the definitions proposed by Nehaniv et al. [8] and no further rules to disambiguate the classification were used (for a review of the categories used see [24]). The observers/coders had the description of the categories and attributed the categories to the behaviors according to their interpretation. Nevertheless, the observers/coders did watch the video of one of the participants together and discussed the classification as a way to train their coding skills and agreement. However, the results regarding the inter-rater agreement were not satisfactory. This lead to the development of a new version of the coding scheme, for which intercoder results are given below.

The second coding scheme also follows Nehaniv et al [8]. In this version, however, the definition of the categories was reformulated and attributes for the categories were defined to facilitate the coding of the video recordings. Furthermore

---

[1]These are not a partition but a covering of instances of gestures, i.e. a given instance of gesture could potentially have more than one of the functions indicated.

---

the following coding heuristics were developed (for a review of the categories used see [23]):

- Eye gaze: only code eye gaze when there is informational value for the interactional gestures category.
- Symbolic gestures: if the episode shows more than one gestural symbolization choose the one you consider more important for the episode and make comments regarding any others. Coding a gesture that performs an action involves the choice between symbolic or manipulative categories. The coder may need to see what the following gesture is and also evaluate to what extent the gestural action is used symbolically from the context.
- Interactional gestures: whenever two similar events follow each other consecutively code as one long episode (i.e. as a single gesture of longer duration).

V. RESULTS OF USER STUDIES

In this section we will summarize some results from two distinct user studies: the evaluation of our coding scheme, descriptive statistics regarding the frequencies and duration of different gestures, types of sequences observed as well as the frequency of co-occurring gestures (distinct gestures overlapping in time).

A. Intercoder Agreement

One of the research aims of the user studies was the development of a reliable coding scheme. Issues with intercoder agreement in the initial user study led to the exclusion of the data from 4 of the 9 participants from this discussion. The second study, however, had a high degree of intercoder agreement which was assessed by Cohen's kappa (Kappa ranging from 59-95 across the functional categories of gestures), suggesting that changes made to the coding scheme after the initial study increased its reliability.

B. The First User Study - Miming

The frequency of gestures produced for each category can be found in table I.

Table I suggests that there were two strategies displayed by the subjects in the first study to teach robots, (1) miming of the shape and use of objects as well as the actual task to be performed and (2) the use of referencing to denote objects and target locations. One of the participants also made use of symbolic gestures.

The duration of the behaviours was also measured. The duration of the majority of referencing behaviours was two seconds or less. The duration of miming manipulation and transportation of objects was mostly two seconds or less, although some instances behaviours having a duration of up two three seconds or less was exhibitited by the participants.

When investigating the sequencing of behaviours, the most frequent sequence was that of referencing object and manipulation followed by referencing place and miming transportation.

Fig. 3 (left) is an example taken from the first user study and it shows the subject pointing to the cutlery with her right hand. The gesture was unambiguously coded as pointing.

TABLE I

Categories	A		Af		B		D		K	
					Participants					
Ref. Objt	78	49.7%								
Ref. Plc	77	48%	7	3.1%						
Miming Trsp.			77	34.2%	73	46.8%	85	48.9%	76	47.8%
Miming Mnp.			76	33.8%	74	47.2%	85	48.9%	77	48.4%
Symbolic Gst.			58	25%						
Gesture Crt.			4	1.4%						
Expressive	1	0.6%			4	2.5%				
Incidental	1	0.6%	3	1.3%	5	3.2%	2	1.1%	6	3.8%

Fig. 3. Examples from the first study – Left: Subject pointing, Right: subject pointing with left hand to general target locus for group of objects and at the same time showing position relative to other objects with the right hand which mimes object transport.

TABLE II

DISTRIBUTION OF GESTURES. DESCRIPTIVE STATISTICS FOR NUMBER OF OCCURENCES FOR ALL CATEGORIES BY EXPERIMENTAL CONDITION

Conditions	Categories	N	Mean	SD	Min.	Max.
Gestures	Pointing	10	.20	.2	0	1
	Interactional	10	7.60	5.15	2	20
	Irrelevant	10	.40	.70	0	2
	Manipulative	10	20.00	9.03	11	39
	Side effect expr.	10	.20	.42	0	1
	Symbolic	10	.20	.42	0	1
Gesture	Pointing	10	1.50	3.48	0	11
and	Interactional	10	5.30	6.80	1	24
speech	Irrelevant	10	.30	.95	0	3
	Manipulative	10	16.20	6.30	8	32
	Side effect expr.	10	1.20	1.75	0	4
	Symbolic	10	.10	.32	0	1

Co-occurrences of gestures was also investigated. The aforementioned issues with intercoder agreement made this difficult. In the included results, mainly one of the participants displayed multiple co-occurrences, which mostly consisted object referencing at the same time as transportation. However, Fig. 3 (right) shows an example of co-occurrence taken from the first user study: the subject is showing the target location to place the fork with his left hand and, at the same time, the right hand is showing that the location is relative to the position of the plate. In this case, the gesture was coded as pointing and miming transportation.

C. The Second User Study: Gesture vs. Speech & Gesture

The descriptive statistics for the number of occurrences for all categories of gestures for both experimental conditions can be found in table II for 10 subjects. Table II suggests that for both conditions, the use of gestures is similar. Manipulative gestures have the highest frequency, followed by interactional gestures. A Wilcoxon test found a significant difference between the use of interactional gestures between the the two conditions, where these gestures were performed significantly more often in the *gestures and speech* condition than in the *gestures only* condition (z=-2.02; p=.045).

Co-occurrences of gestures were also found in the second study, the majority of which were interactional gestures co-occurring with manipulative gestures.

The frequencies for the different time intervals for all categories gestures according to experimental condition can be found in table III and the descriptives for durations can be found in table IV. Table III suggests that the majority of time intervals for all categories apart from manipulative

lies under 2 seconds in both conditions. For the manipulative category, however, this is not the case. For this category the majority of time intervals lies above 3 seconds. The same pattern is confirmed by the durations reported in table IV. For all categories apart from manipulative, the mean duration was below 2 seconds. For manipulative gestures, the mean for both conditions was 3.3 seconds.

Fig. 4. Examples from the second study: Manipulative gesture - grasping a cup (left), symbolic gesture - index finger raised introducing the FIRST STEP (right)

Fig. 4 (left) was taken from the second user study. We can see the subject manipulating the cup in order to place it just in front of the plate. In this the coders classified the gesture as manipulative. Fig. 4 (right) shows one subject from the second user study displaying a symbolic gesture. More specifically, the subject intended to communicate that he was about to show step one of his explanation.

It is important to note, however, that the duration of the manipulative gestures was related to the physical layout of the scenario. Some of these gestures required the transporta-

TABLE III
FREQUENCIES FOR DIFFERENT TIME INTERVALS BY EXPERIMENTAL CONDITION

Conditions	Categories	Time Intervals in seconds					
		< 1	1-2	2-3	3-4	4-5	5-6
Gestures	Pointing	2	-	-	-	-	-
	Interactional	4	6	-	-	-	-
	Irrelevant	2	1	-	-	-	-
	Manipulative	-	2	3	1	3	1
	Side effect expr.	1	2	-	-	-	-
	Symbolic	1	1	-	-	-	-
Gestures and Speech	Pointing	2	1	-	-	-	-
	Interactional	1	7	1	1	-	-
	Irrelevant	1	-	-	-	-	-
	Manipulative	-	-	5	2	2	1
	Side effect expr.	1	1	-	-	-	-
	Symbolic	-	1	-	-	-	-

TABLE IV
DURATION OF GESTURAL CLASSES EXHIBITED BY THE SUBJECTS IN THE TWO EXPERIMENTAL CONDITIONS. DESCRIPTIVE STATISTICS IN SECONDS FOR ALL CATEGORIES BY EXPERIMENTAL CONDITION

Conditions	Categories	N	Mean	SD	Min.	Max.
Gestures	Pointing	2	.32	.05	.28	.36
	Interactional	10	1.21	.50	.53	1.98
	Irrelevant	3	1.37	1.09	0.6	2.62
	Manipulative	10	3.28	1.31	1.16.	5.18
	Side effect expr.	2	.62	.37	0.36	.89
	Symbolic	2	1.08	.34	.84	1.32
Gestures and Speech	Pointing	3	.79	.42	.36	1.19
	Interactional	10	1.63	.70	.81	3.30
	Irrelevant	1	.13	.	.13	.13
	Manipulative	10	3.30	.99	2.10	5.00
	Side effect expr.	4	1.38	.64	.85	2.28
	Symbolic	1	.14	.	.14	.14

tion of objects to the other side of the table, a task that sometimes was solved by the participant walking around the table while performing the task. This is reflected in the comparatively large standard deviation for the mean duration of this task. This suggests that any time window for the capture of gestures needs to be flexible to take into account the physical parameters of a task.

In the first user study, participants were asked if they would like to learn a set of predefined gestures that could be used in instructing a robot to perform a variety of tasks, or if they would prefer to teach the robots the gestures they would like to use with it. The sample was split into two equal groups, where roughly half preferred a set of predefined gestures while the other half argued for a system that would be able to learn gestures from the user. Almost all participants, however, did state that they would be willing to learn a set of predefined gestures from a manual, if it was necessary to interact effectively with the robot.

The second user study had a stronger focus on the perceived differences between speech and gestures. All 10 participants indicated that they would prefer a combination of speech and gestures to communicate with a robot. When asked about the possibility of using a set of predefined gestures or words, 8 of the 10 participants indicated that they would be willing to learn a set of gestures and words to interact with the robot.

VI. DISCUSSION OF THE OBTAINED RESULTS

An activity recognition system which makes use of sensors which are onboard the robot will always be limited in the granularity, variety and number of observable movements and activities. Thus, it important to focus the recognition system on those movements and gestures which are important for the HRI context.

The aim of this work is twofold: On the one hand, the recognition process must be optimized and adapted according to the requirements and characteristics provided by the user study results. On the other hand, the user studies evaluate to what extent people are willing to adapt to a recognition system in order to communicate with the robot.

Several tentative conclusions can be drawn that could serve to inform the future design and adaptation of gesture recognition systems, at least for those used in contexts similar to those in the experimental scenarios:

- According to Tables III and IV, the duration of gestures in nearly all classes can be assumed to be roughly between 0.5 and 3 seconds. This has strong effects on the design of a recognition system: The time window for feature extraction needs a maximum size slightly more than 3 seconds, and a time-dependent classifier only has to consider this time frame. Future studies may show whether the duration can even be used as validation for gesture recognition.
  The only exception concerns manipulation, which has longer durations occurring. This case has to be considered seperately for recognition, which is simple if also extrinsic features are taken into account.
- The user studies have shown that certain gestures often appear in a sequence. This is e.g. the case for *referencing object* and *manipulation*. This fact can be used in a classifier to recognize gestures and activities. E.g. using a *Bayesian classifier*, these relations can be modeled and improve recognition.
- According to sec. V-B, certain gestures and activities may occur in parallel. It is important to note that this also entails that other combinations of activities can *not* occur in parallel. This is very important for classification and interpretation of the resulting feature and activity set: Conflicts (false parallel classifications) can be detected and solved, e.g. using a rule-based system, or by directly modeling these relations in the classifier's activity model.
- From the user study evaluations, we learn that people are to some extent willing to adapt to the recognition system and to learn new gestures for communication with a robot. This gives major implications for the development of the HRI system: It is not necessary to develop a perfectly human-like gesture recognition. In contrary, one has to find the compromise between recognition capabilities, recognition efficiency/robustness and willingness to adapt to the system. Following the user study evaluation it is more important to combine different modalities like gesture recognition and speech.

- Table I shows that, even if the different gesture classes occur in similar frequencies across all subjects, the variations within the demonstrations of one subject are very small. This indicates that a single person tends to use only a subset of all gestures or activity classes. For the recognition process, this leads to a *personalization* of the classification and interpretation process. Depending on the currently observed person, only a subset of all activities must be detected.

Determining an optimal set of gestures for a given HRI context which is necessary to recognize in order to obtain an efficient, general purpose, but still robust communication and interaction modality is an open issue. Such sets must be determined in further user studies. The main challenge is to find a set of gestures which people are willing to use, which enables efficient communication, and which is observable with the robot's onboard sensors and capabilities. The aim of the next study towards achieving this is to develop an understanding for the features of gestures that people are willing to perform, and of those which people would not accept. It will also serve to inform the design of the system feedback during human-robot interaction, e.g. when to ask the user to repeat his gesture, or when and how to acknowledge a recongized gesture.

VII. CONCLUSION

This paper has presented an approach for human activity recognition. It utilizes a 3D human body tracking to continuously generate features. These features are used as input for a neural net classifier, which has been trained with a set of activity examples.

In parallel, two user studies have been carried out to gain knowledge about characteristics, preferences and requirements for gestures in a human-robot teaching scenario. The results of these studies have been evaluated with regard to human activity recognition: Performed gestures have been analyzed to obtain information on gesture sequences, parallel execution, frequency of occurence etc. This can in turn be included in the activity model for the recognition process.

One important result is that it is not necessary to imitate the full human capabilities for gesture recognition. People are to some extent willing to adapt to a system's limitations. Finding this optimum between unconstrained gesturing, adaptation to the robot's capabilities and system robustness will be part of our future work.

REFERENCES

[1] K. Dautenhahn, "The art of designing socially intelligent agents: Science, fiction, and the human in the loop," *Applied Artificial Intelligence*, vol. 12, pp. 573–617, 1998.

[2] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, p. 143, 2003.

[3] J. Cassell, "Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, Massachusetts: The MIT Press, 2000, pp. 1–27.

[4] A. Kendon, "Gesture," *Annual Review Of Anthropology*, vol. 26, pp. 109–128, 1997.

[5] D. McNeil, *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago, 1992.

[6] D. McNeill, *Gesture and Thought*. Cambridge University Press, 2005.

[7] S. Goldin-Meadow and S. M. Wagner, "How our hands help us learn," *Trends In Cognitive Sciences*, vol. 9, no. 5, pp. 234–241, 2005.

[8] C. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami, "A methodological approach relating the calssification of gesture to identification of human intent in the context of human-robot interaction," in *Proc. IEEE Ro-Man*, 2005, pp. 371–377.

[9] S. Vacek, S. Knoop, and R. Dillmann, "Classifying human activities in household environments," in *Workshop at the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[10] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *Proc. IEEE Intl. Conf. Robotics and Automation (ICRA)*, Orlando, Florida, 2006.

[11] S. Knoop, S. Vacek and R. Dillmann, "Modeling Joint Constraints for an Articulated 3D Human Body Model with Artificial Correspondences in ICP," in *Proceedings of the International Conference on Humanoid Robots (Humanoids 2005)*. Tsukuba, Japan: IEEE-RAS, 2005.

[12] C. Cédras and M. Shah, "Motion-based recognition: a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, March 1995.

[13] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, January 1999.

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, March 1999.

[15] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Reconition*, vol. 36, no. 3, pp. 585–601, 2003.

[16] P. C. Ribeiro and J. Santos-Victor, "Human activity recognition from video: Modeling, feature selection and classification architecture," in *Proceedings of the International Workshop on Human Activity Recognition and Modelling 2005*, vol. 1, 2005, pp. 61–78.

[17] F. Pla, P. Ribeiro, J. Santos-Victor, and A. Bernardino, "Extracting motion features for visual human activity representation," in *Pattern Recognition and Image Analysis: Second Iberian Conference, IbPRIA 2005, Estoril, Portugal, June 7-9, 2005, Proceedings, Part I*, J. S. Marques, N. P. de la Blanca, and P. Pina, Eds., vol. 3522. Springer, 2005.

[18] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Segmentation and classification of human activities," in *Proc. HAREM International Workshop on Human Activity Recognition and Modelling*, 2005.

[19] A. K. Dey, "Understanding and using context," in *Proceedings of Personal and Ubiquitous Computing*, 2001.

[20] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[21] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[22] N. Kwak and C.-H. Choi, "Improved mutual information feature selecor for neural networks in supervised learning," in *Proc. International Joint Conference on Neural Networks, 1999*, vol. 2, 1999, pp. 1313–1318.

[23] N. Otero, C. Nehaniv, D. Syrdal, and K. Dautenhahn, "Naturally occurring gestures in a human-robot teaching scenario: an exploratory study comparing the use of gestures only or gestures and speech," in *Proc. IEEE Ro-man 2006 (this volume)*, 2006.

[24] N. Otero, C. Nehaniv, K. Dautenhahn, J. Saunders, and A. Alissandrakis, "Naturally occuring gestures in a human-robot teaching scenario: An exploratory study," School of Computer Science, University of Hertfordshire, Tech. Rep. 443, 2005.