Abstract—Human activity recognition is an essential ability for service robots and other robotic systems which are in interaction with human beings. To be proactive, the system must be able to evaluate the current state of the user it is dealing with. Also future surveillance systems will benefit from robust activity recognition if realtime constraints are met, allowing to automate tasks that have to be fulfilled by humans yet.

In this paper, a thorough analysis of features and classifiers aimed at human activity recognition is presented. Based on a set of 10 activities, the use of different feature selection algorithms is evaluated, as well as the results different classifiers (SVMs, Neural Networks, Bayesian Classifiers) provide in this context. Also the interdependency between feature selection method and chosen classifier is investigated. Furthermore, the optimal number of features to be used for an activity is examined.

I. INTRODUCTION

Human activity recognition is an important ability in the field of human robot interaction (HRI). Besides other channels like perception of environment, internal state, context and system targets, it is one of the essential informations. A personal service robot must be capable of detecting and recognizing human activities in order to decide about next actions. To offer help to a person, the robot has to detect and understand the user’s intentions and to infer his goals. Especially when confronted with humans who are not familiar with service robots and their behavior, the system must detect if the person needs help or guidance and approach him in a proper way. It is even possible that the robot should not interrupt a user in his current activity. But again it is necessary for the robot to recognize the activities of the human to make a correct decision. During interaction with a human, it is imperative to understand human activities, as important information is conveyed over this communication channel.

Within the project COGNIRON, components and methods for building a robot companion are developed. A robot companion is defined as a robotic system, able to assist humans in their daily life at home and learning continuously to improve its abilities to serve. Such a robot must be able to perceive its environment and the people around it and to learn simple and more complex tasks. To be a real companion for its user, it must be proactive, offering helpful services when possible.

Our system for activity recognition system is developed in this project and uses other systems as input. One basis is a subproject concerned with observation and tracking of human body motion and body configuration. The result of this tracking is used as one input channel for the activity recognition process. Detection and recognition of objects, which are conducted in another part of the project, can serve as other input channels. Previous work [1] showed that our approach for the activity recognition, described in sec. III, is feasible. This paper addresses the problem of activity recognition and feature selection, in particular finding an optimal set of features for a given set of activities. These activities are chosen to cover a typical scenario for a robot. Furthermore, optimal classifiers are investigated. Additionally, the dependencies between classifiers and feature selection methods are explored.

The paper is organized as follows: Section II gives an overview on current research in the field of activity recognition. In section III, a short overview of our system for activity recognition is given. Furthermore, detailed information about the activity set as well as the base feature set is given. The following section IV presents our approach for the evaluation and comparison between different feature sets and classifiers. Subsequently, the results of the experiments will be discussed in section V. The work closes with conclusions and a short prospect of future work in section VI.

II. STATE OF THE ART

Activity recognition is a very active research field. Good starting points to the general area are the surveys in [2], [3] and [4].

Different approaches exist for the problem of classifying human activities, involving classifier types of every kind. Exemplarily, usage of hidden markov models in the context of human activity recognition can be found in [5] and [6]. Neural Networks are used in [7], while [8] uses bayesian networks to classify human activities.

On a higher level of abstraction than addressed in this paper, sequences of simple activities are combined to a more global view of the activities, e.g. to distinguish between talk and question phase in a lecture situation. Therefore, the advanced utilization of context information is explored, for instance in [9], [10] and [11].

Although all these approaches show promising results, usually no justification is given for the usage of one classifier or the other. Furthermore, the method how reasonable
features for the recognition are chosen is not compared to any other method, so again a justification for the choice is not given. To close this gap, we are going to evaluate different classifiers (SVMs, Neural Networks, Bayesian Classifiers) against each other in order to decide about the optimal component for the problem of human activity recognition in our setup. The decision about an appropriate method for feature selection will also be discussed.

III. BACKGROUND

The work presented in this paper aims at typical home scenarios where a proactive robot could offer services. A set of 10 different activities has been derived from such a typical home scenario placed in a living room, table I shows the list of activities. The following is an example for a typical sequence formed by these activities: The observed person sits down and then either reads a book or types on a laptop, leading to the activities (3), (4) and (5) in tab. I. More information about the selection of these activities can be found in [12]. As the chosen activity set contains activities involving very different movements, we assume the results to generalize very well on other household scenarios as well.

TABLE I
LIST OF ACTIVITIES IN THE KEY EXPERIMENTS

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>STANDING</td>
</tr>
<tr>
<td>(2)</td>
<td>PUT OBJECT ON TABLE</td>
</tr>
<tr>
<td>(3)</td>
<td>SITTING</td>
</tr>
<tr>
<td>(4)</td>
<td>READ A BOOK</td>
</tr>
<tr>
<td>(5)</td>
<td>TYPE ON A LAPTOP</td>
</tr>
<tr>
<td>(6)</td>
<td>PULL HAND BACK</td>
</tr>
<tr>
<td>(7)</td>
<td>HOLD OUT BAND</td>
</tr>
<tr>
<td>(8)</td>
<td>PULL OBJECT BACK</td>
</tr>
<tr>
<td>(9)</td>
<td>HOLD OUT OBJECT</td>
</tr>
<tr>
<td>(10)</td>
<td>TAKE OBJECT FROM TABLE</td>
</tr>
</tbody>
</table>

The approach for the activity recognition cycle consists of the following steps:
- Tracking (+ Observation of environment)
- Feature selection
- Classification

The human is tracked and the tracking results serve as input features, together with features which are provided by other modules such as an object recognition. In the next step, out of the raw set of features, a subset is selected, which contains the most relevant features for the recognition. The features within this subset are used in the classification step. The features are used as input for trained classifiers (one for each activity), which decide whether the observed person is engaged in one or more known activities. A detailed description of the approach for activity recognition can be found in [1].

Basis for the selection of feature subsets is a raw feature set, that contains 323 different features for each frame. Table II gives an overview on all kinds of features which are available, a complete list of all features can be found in [12]. These features were chosen considering the limits of the tracking (i.e. which features can be extracted at all), and using the results of user studies (details can be found in [13]).

IV. EVALUATION APPROACH

This paper presents our insights of a comparison of feature selection methods and classifiers for recognizing human activities. The optimal number of activities covered by a single classifier is examined in sec. IV-A. Another interesting aspect is the correlation between the feature selection method and the type of classifier. Also, the correlation between optimal classifier and observed set of activities will be investigated. In general, an analysis of different classifiers, their absolute performance and their relatively compared performance, helps to build a robust and powerful system. Sections IV-B to IV-E present our approach to examine these issues.

A. Granularity of the classifier

A single classifier can cover all activities, a group of activities or a single activity. Using one classifier for all activities is unfavorable because it is difficult to detect two concurrent activities such as e.g. standing and waving at the same time with this technique. Covering a group of activities by a single classifier has the drawback that all groups have to be defined manually which is infeasible most of the time. Using a classifier for each single activity independently has the disadvantage of requiring a big amount of training while background knowledge about mutually exclusive activities cannot be applied (e.g. the mutual exclusivity of SITTING, STANDING, WALKING can not be exploited directly). However, it is modular as it supports the exchange of activities without complete retraining. Comparing the characteristics of these techniques, using one classifier for each activity is most preferable in practice. So for the work presented in this paper, the term classifier applies always to a classifier trained for the distinction between two classes, the presence or absence of the trained activity.

B. Test data acquisition

In order to compare different classifiers for the recognition of the defined set of 10 activities, a number of sequences had to be recorded as input for classifier training and evaluation. For each activity, 12 sequences have been recorded and segmented by hand, each with an average length of 75 frames.
For the sake of clarity, only right hand gestures have been analyzed. Although the left hand could be incorporated easily as well, it was not utilized in our experiments as the use of mirrored actions with left and right hand does not add much insight into the feature selection and classification process.

In the following, both stages of the approach, the feature selection and the classification based on the selection results, are presented.

C. Feature Selection

In the first stage, two different algorithms for selecting relevant features have been applied to each of the activities.

- **Evaluation of Information Gain from Attribute (EIGA):** This algorithm is based on the information theoretical concepts of information content and entropy. The utility of a feature is evaluated by measuring the information gain with respect to the activity class.
- **Correlation-based Feature Subset Selection (CbFSS):** This algorithm evaluates the utility of a subset of features by considering the individual predictive ability of each feature (comparable to the information gain by using the feature) along with the degree of redundancy between them.

The resulting sets of selected features have both been used in the second stage. The results have been compared to get evidence which of these methods should be chosen to be used for future work.

D. Training and classification

In the second stage, an increasing number of features from the preselected feature set was used to train and test different classifiers in order to decide about a) the optimal number of features, and b) the optimal classifier for recognition. Up to 10 features were used with the following classifiers:

- Naive Bayes Network (NBN)
- Bayes Network (BN)
- Multilayer Perceptron (MLP)
- Radial Basis Function Network (RBF)
- Support Vector Machine (SVM)

Different kernel functions for the SVMs have been tested with different parameter settings. The tested kernels include homogeneous and inhomogeneous polynomial kernels of different order as well as RBF kernels with different parameter settings. Of the different SVM variants which were tested, the SVM with RBF kernel performed best, hence only results of this SVM will be shown in the following presentation of results.

E. Test procedure

For each of the defined activities, the complete input data was used in the feature selection steps for both methods (EIGA and CbFSS), resulting in two feature sets for each activity.

The input data\(^1\) has then been prepared as follows for each of the activities: It was randomly split with \(\frac{2}{3}\) used as training data and \(\frac{1}{3}\) as test data. This action was performed with feature set sizes from 1 to 10 for both selected feature subsets. The prepared data was then used to train and evaluate each classifier, fig. 1 depicts the approach.

Splitting, training and evaluation was repeated 10 times, to avoid random effects from the splitting. The results are presented and analyzed in the next section.

V. RESULTS

The base feature set, extracted from the output of the perception components, contains absolute positions and velocities of the observed human. These absolute features have been removed for the following experiments (except for the torso velocity), only the positions and velocities relative to the rest of the body have been utilized. This was done to avoid the selection of features which depend strongly on sensor position or coordinate system origin. As the set of training data is of course limited, this is a probable issue. The correctness of this idea was proven by some test runs with the complete feature set. Exemplarily, both feature selection strategies selected three absolute limb positions for the activity READ A BOOK.

A. Results of the feature selection

The results of the selection in the revised feature set were 10 features for each activity of the defined set. In table III, the first 5 features selected by CbFSS for each activity are shown. These (and the features 6–10, not shown here) are the features which were used for the following comparison between different classifiers and the evaluation of the two feature selection methods.

Most of the selected features are well chosen and also intuitive from a human perspective, see for example the chosen features for the activity STANDING. The first chosen features consist of the **absolute angle of left and right thigh** (CbFSS) and the angle between left lower leg and left thigh and the **position of one foot relative to thigh and relative to torso** (EIGA) respectively. All these features describe observations which a human uses to recognize STANDING: upper legs are vertical, angle between lower leg and thigh is near 180°, and feet are positioned below torso and thigh. In the chosen features for the activity READ A BOOK, the first three features are the same for both selection methods (although not chosen in the same order): **vicinity of hands to a book**, the angle between left upper arm and torso, and

---

\(^1\)The training data is available in the ARFF format via [http://www.iaim.ira.uka.de/users/loesch/dataset10activities.tar.gz](http://www.iaim.ira.uka.de/users/loesch/dataset10activities.tar.gz)
TABLE III
FIRST 5 FEATURES SELECTED BY CBFSS FOR EACH ACTIVITY.

<table>
<thead>
<tr>
<th>STANDING</th>
<th>PUT OBJECT ON TABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Abs. ang., left thigh</td>
<td>Abs. ang., right upper arm</td>
</tr>
<tr>
<td>2. Abs. ang., left thigh</td>
<td>Abs. ang., right thigh</td>
</tr>
<tr>
<td>3. Abs. ang., right thigh</td>
<td>Ang., left thigh → torso</td>
</tr>
<tr>
<td>4. Ang., left lower leg → thigh</td>
<td>Pos., right foot → thigh</td>
</tr>
<tr>
<td>5. Ang., right lower leg → thigh</td>
<td>Abs. ang., right lower leg</td>
</tr>
</tbody>
</table>

SITTING

<table>
<thead>
<tr>
<th></th>
<th>READ A BOOK</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Abs. ang., left thigh</td>
<td>Distance, left hand ←→ right hand</td>
</tr>
<tr>
<td>3. Abs. ang., right thigh</td>
<td>Ang., left upper arm ←→ torso</td>
</tr>
<tr>
<td>4. Ang., left lower leg → thigh</td>
<td>Ang., left thigh → torso</td>
</tr>
<tr>
<td>5. Ang., right lower leg → thigh</td>
<td>Vicinity, hand ←→ object</td>
</tr>
</tbody>
</table>

TYPE ON A LAPTOP

<table>
<thead>
<tr>
<th></th>
<th>FULL HAND BACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vicinity, hand ←→ laptop</td>
<td>Ang., left upper arm ←→ torso</td>
</tr>
<tr>
<td>2. Abs. ang., left forearm</td>
<td>Ang. vel., right upper arm ←→ torso</td>
</tr>
<tr>
<td>3. Mean pos., left hand</td>
<td>Abs. ang., right thigh</td>
</tr>
<tr>
<td>4. Abs. ang., torso</td>
<td>Vicinity, hand ←→ object</td>
</tr>
<tr>
<td>5. Abs. ang., right thigh</td>
<td>Abs. vel., torso</td>
</tr>
</tbody>
</table>

HOLD OUT HAND

<table>
<thead>
<tr>
<th></th>
<th>FULL OBJECT BACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ang., left upper arm → torso</td>
<td>Ang. vel., right upper arm ←→ torso</td>
</tr>
<tr>
<td>2. Abs. ang., torso</td>
<td>Vicinity, hand ←→ right upper arm</td>
</tr>
<tr>
<td>3. Vicinity, hand ←→ object</td>
<td>Mean pos., right foot</td>
</tr>
<tr>
<td>4. Ang., right upper arm → torso</td>
<td>Abs. ang., right lower leg</td>
</tr>
<tr>
<td>5. Mean pos., left hand</td>
<td>Abs. vel., torso</td>
</tr>
<tr>
<td>5. Covariance, left hip ←→ right hip</td>
<td>Ang., left thigh → torso</td>
</tr>
</tbody>
</table>

distance between left and right hand. Again, all these features would be chosen by a human manually as well.

But not all of the chosen features are reasonable to human sense. For example, the features chosen for the activity HOLD OUT OBJECT contain mean position of right foot and absolute angles of thighs, while vicinity of hand to object is chosen only by one of the two selection methods, and even there it does not gain a high relevance measure. The reason for this incidence and a possible solution are explained in section V-E.

B. Comparison between feature selection methods

Both feature selection methods have been used for training and classification on the test data set. While in the usage with NBNs and a very small number of features (1 or 2) EIGA performs better than CbFSS, the latter shows its strength in the use of more features and more elaborated classifier types. The gap between both methods widens even more if more features are used. But even then, the difference of the recognition rates is in the small one-figure percent range.

Altogether, the experimental results show that CbFSS is usually slightly better than EIGA, independent from the chosen classifier. The difference averages a 1% better recognition rate by the use of CbFSS for the feature selection. Therefore, for further research only CbFSS will be used. Accordingly, results of the experiments presentend in the following sections are obtained by using features which were selected exclusively by CbFSS.

C. Correlation between classification quality and number of features

Each classifier has been applied to all activities, with an increasing number of features. Fig. 2 shows exemplarily the results yielded by the SVM. The other classifiers showed similar results, which will be discussed in the following as well.

The activities READ A BOOK, SITTING, STANDING and TYPE ON LAPTOP are omitted in further discussion because they are classified perfectly by all classifiers. But HOLD OUT HAND, HOLD OUT OBJECT and TAKE OBJECT FROM TABLE allow to reason about the optimal number of features for classification. They converge to a stable (but not perfect) recognition rate which only grows very slowly with an increasing number of features.

NBNs show stable recognition rates with three or more features for the mentioned activities. But even more noticeable, NBNs are apparently sensible against bad or noisy features. This effect is demonstrated in the activity PUT OBJECT ON TABLE. Recognition rate for this activity has a maximum using 2 features, but drops then until a minimum is reached using 6 or 7 features.

BNs, MLPs, RBFs and SVMs reach a stable recognition rate with 3 or 4 features. The only exception is shown by the combination of MLPs and the activity TAKE OBJECT FROM TABLE, which converges with 2 features already. All these classifiers are not as sensitive against bad features as NBNs, showing monotonic growth for all activities.

The correlation between number of utilized features and classification quality differs strongly between activities even for the same classifier. An example for this fact are the activities PULL BACK HAND and TAKE OBJECT FROM TABLE, the former being recognized with over 95% correctness, while the latter hardly reaches 90% (with BN, RBF, SVM) or even stays below (NBN, MLP). For most activities, the classification quality is nearly constant over the number of features. Some activities show improved classification with several classifiers when using more than three features. Generally, the gain for using 10 instead of 5 feature is negligible.

D. Comparison of classifiers

Fig. 3 shows the average over all activities’ recognition rate of all classifiers. As can be seen easily, the NBN classifier performs clearly worse than all other classifiers. Furthermore, it does not have any advantage of a larger features subset, having its peak with 6 features and performing even worse with 10 features than with 7 or 8 features. The other classifiers are rather close, but RBF performs slightly worse than the other classifiers using a small number of
features and does not take advantage from the use of 5 or more features as the other classifiers do.

Fig. 3. Comparison between average classifier results over all activities

To narrow the view, both NBN and RBF are discarded and further analysis will concentrate on the performance of BN, MLP and SVM. In fig. 4, the recognition rates of 6 activities which show significant differences between the classifiers are shown for the three classifiers using 3 – 6 features. The remaining activities do not reveal much information because of their lack of distinction between the classifiers, showing equal recognition rates, and are therefore omitted.

The remaining 6 activities show apparent but not large differences between the three classifiers. Particularly, there is not one classifier outperforming the others. BNs perform the recognition of HOLD OUT HAND (fig. 4, 1st row, left) on the same high level independent of the number of features, superior to MLP and SVM with less than 6 features. Also, the activities HOLD OUT OBJECT and PUT OBJECT ON TABLE are recognized better by BNs than by MLP or SVM. The major downside of BNs compared to MLP and SVM is apparently that the recognition rate does not increase much with the number of used features. Another drawback of BNs is observable in the diagram of PULL HAND BACK (fig. 4, 2nd row, left): it is possible that the recognition rate decreases with a growing number of features. Aside from that, BNs perform as good as MLPs and SVMs and produce results of about 95% for most activities, and even a bad performance has rates over 85%.

MLP and SVM show the same behavior and very similar recognition rates throughout all activities. They both produce results between 90 and 95% for most activities, and even for PULL HAND BACK, which yields the worst results within this evaluation, the rates are still over 85%. Overall, both tend to increase their recognition rate with increasing number of features, observable e.g. in the diagram of TAKE OBJECT FROM TABLE (diagram in fig. 4, 2nd row, right). This activity is also an example for MLP and SVM performing noticeably better than BNs.

Considering all these facts and the very high recognition rates produced by BNs as well as MLPs and SVMs leads to the conclusion, that the information about the current activities is coded very clearly in significant features. Hence it is not possible or necessary to pick out one single optimal classifier. In our framework, 4 features and SVMs are used for the activity recognition as they perform as well as MLPs, but do scale better than BNs with the number of selected features.

A possible field of future investigations is the differing behavior of SVM and BN. It seems reasonable to train both classifiers concurrently and evaluate them against each other after the training to select the appropriate classifier on a per activity basis.

E. Lessons learned

Some interesting issues can be analyzed beyond noticing the generally high quality of classification results. In fact, the similarly strong performance of all classifiers with quite distinct underlying techniques leads to the conclusion that an individual feature has a high intrinsic information value. Therefore, further considerations about the choice of classifiers are mostly superfluous. On the other hand, careful selection of features is important to improve the quality of activity recognition even more.

A closer look into actual feature selection gives insight for further optimization of the process. For some activities, features are selected which are in considerable distance to the main spatial region of the activity. So e.g. for some activities which are confined to the upper body, features of the lower body are selected and vice versa. Fig. 5 highlights this issue on the classification of three activities, marking the chosen features (whereas circles indicate positions, angles and periodicities, arrows indicate velocities, and lines indicate distances).
One way to realize the improvement would be the application of a much larger training set, especially with very similar positive and negative examples. This way, a very specific and yet robust feature selection could be achieved. Unfortunately, this approach produces new problems, as gathering a large enough example set is practically infeasible for casual interactive learning of a robot companion.

On the other hand, the interactivity can be exploited and coupled with the use of background knowledge. Rather than conducting a large number of demonstrations, a human instructor can point the robot explicitly towards regions of special interest. So while performing the activity during a demonstration, the instructor leads the focus of the observing robot towards the regions which include the most important features. This can be done by gestures and supporting speech. Applying research results from [14] and using the instructor’s hints, the robot can exclude irrelevant features and retain only features from the body parts pointed out. This approach reduces the number of irrelevant features strongly without much additional effort for the instructor. Background knowledge can be used to deal with mutually exclusive activities. The choice of using an individual classifier for each activity already permits explicit handling of the attribute of two activities being mutually exclusive as discussed in section IV-A. Yet, to exploit this opportunity in depth, a rule-based system has to be applied which expresses and processes mutually exclusive relations. On top of the classification, this system applies model knowledge about the relations to verify the classification by detecting and solving inconsistencies.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, the use of different feature selection methods and different classifiers in the context of human activity recognition has been investigated. We found that a number of 4 features is sufficient to have a robust recognition, while the gain of using more features is negligible. We figured out that there is no single optimal classifier for our purpose. In fact, the information about current activities is apparently coded very clearly in significant features. In our framework, SVMs are used for the activity recognition. SVMs produce recognition rates of at least 87% and an average rate of about 95% in the evaluated activities.

In the course of this evaluation, we also found that features selected with CbFSS give better results than those selected by EIGA, independent of the classifier actually used. Hence, CbFSS is used in our framework for the selection step.

At last, we have presented our perspective for a stronger incorporation of the user into the learning process of his robot companion by using hints from the user to correct the feature selection.

VII. ACKNOWLEDGMENTS

The work described in this paper was conducted within the EU Integrated Project COGNIRON (‘The Cognitive Robot Companion’ – www.cogniron.org) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

REFERENCES