The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems

Lukas Rybok*, Simon Friedberger*, Uwe D. Hanebeck*, and Rainer Stiefelhagen*†
* Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
† Fraunhofer IOSB, Karlsruhe, Germany
{rybok, friedberger, rainer.stiefelhagen}@kit.edu, uwe.hanebeck@ieee.org

Abstract—**Human action and activity recognition from videos has attracted an increasing number of researchers in recent years. However, most of the works aim at multimedia retrieval and surveillance applications, but rarely at humanoid household robots, even though the robotic perception of human activities would allow a more natural human-robot interaction (HRI). To encourage future studies in this domain, we present in this work a novel data set specifically designed for the application in HRI scenarios. This Robo-kitchen data set consists of 14 typical kitchen activities recorded in two different stereo-camera setups, and each performed by 17 subjects. To establish a baseline for future work, we extend a state-of-the-art action recognition method to be applicable on the activity classification problem and evaluate it on the Robo-kitchen data set showing promising results.**

I. Introduction

Automatic situation understanding from video is important for many applications, such as video retrieval and surveillance and thus, gains an increasing amount of attention in computer-vision research. Humanoid service robots can also greatly benefit from understanding household situations, since it allows to establish more natural human robot interfaces. As a possible application, the robot may take the role of a butler observing the scene from a point in the background and offering unsolicited help whenever he assesses that it might be required. In this paper we present a data set specifically designed for this scenario since to the best of our knowledge no such data is publicly available yet. Instead, novel approaches in action and activity classification are currently only developed to perform well on benchmarks posing different challenges than those emerging in the humanoid robots domain.

Both terms, action and activity, are used interchangeably in the literature. In the following, we will be using the taxonomy as used by Moeslund et al. [1], where an action describes an *atomic* meaningful motion event, such as *Pick Up Object* or *Jump*. However, activities denote complex action sequences often getting their meaning from interaction with objects or the overall context. It should be noted, that the distinction between actions and activities is not always clear. For instance, the activity *Sweep Floor* might also be regarded as a periodic action due to its quite simple nature.

The aim of our paper is to establish a basis for the comparison of holistic activity recognition approaches for humanoid robots. We hope that our Robo-kitchen data set will encourage researchers to focus their work on this challenging task that is closely related to real-world applications. As a setting for our recordings we selected a kitchen scenario, since it provides a vast range of possible activities for which a robot might offer his help. The setup has been designed to resemble one of a humanoid robot as closely as possible. All of this poses many challenges for view-based activity recognition approaches, such as cluttered background, difficult lighting conditions, (self-)occlusions, different view-points, and a limited field of view. Also, because only a limited amount of space is available on a robotic platform, we only used cameras with small enough optics that makes them capable of being integrated on a robotic head, but also induces noisy video data. Most importantly, we barely restricted the way how the recorded subjects had to perform the activities resulting in a collection of natural motions with much variation as opposed to most currently publicly available data sets. Imitating humanoid robots in our setup however also results in the use of stereo cameras, which can be beneficial for activity recognition, since it allows for person tracking and extraction of motion trajectories in 3D. We argue that while in motion, the robot is already occupied with performing a task and thus it is sufficient for the robot to assess the room situation while standing still. For this reason, only static cameras were used for the recordings.

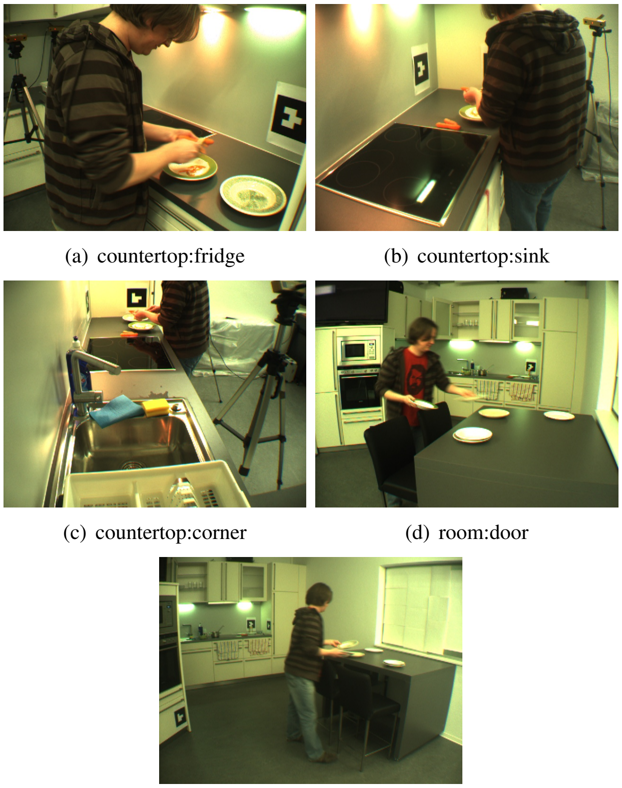
The second contribution of this paper is an extension of a view-based state-of-the-art action classification algorithm to make it applicable for activity recognition. The aim is to provide baseline results for future developments in the area of activity recognition for humanoid robots with the focus on a shortest possible response time.

II. Related work

In recent years, human action and activity recognition from video has gotten much attention in the computer vision community. A detailed overview over the current state-of-the-art can be found in the surveys from Moeslund et al. [1], Turanga et al. [2], and Poppe [3]. To establish a basis for the comparison of novel action recognition approaches, many data sets have been published, which are targeted at different applications. Two very early data sets that have become the de-facto standard action recognition benchmarks are the KTH [4] and the Weizman data sets [5]. Both contain only few and relatively simple, periodic actions, such as *Running* or *Boxing* that are performed in very constrained environments and do not contain much intra-class variations. The IXMAS [6] and the HumanEva [7] data sets are conceptually similar to the aforementioned ones in the sense of a simplified setting. However, they were recorded using a multiple camera setup making the data also suitable to evaluate approaches aiming at view-independent action recognition, which is still a very challenging topic. All of these data sets are of limited relevance to practical applications, since the contained actions are composed of distinct movements often making them appear unnatural. Also, the recorded actions have a lack of variability in body postures when being compared to the same actions performed in the context of daily living activities. Thus, these shortcomings prompted the development of datasets containing more natural and complex actions.

Because it is difficult for people to act naturally when participating in an artificially set data collection, Laptev et al. began with creating data sets based on actions performed in movies [8]–[10]. A similar approach was taken by Liu et al., who collected a set of realistic and very diverse videos from YouTube [11]. This new generation of action data sets contains interaction events between humans, actions involving object manipulation and much variability with respect to viewing angles, lighting, background and actors. Yet, even though the setting got much more challenging, the data sets still contain only very simple, *atomic* actions and are mostly aimed at video retrieval applications.

Recently, many data sets for the evaluation of more complex action sequences, i.e., activities, have been published [12]–[15], all of them set in a kitchen scenario. Besides of video recordings, these data sets also contain motion capture information and data captured with different other sensors (e.g., RFID, microphones, accelerometers). Thus, their primary aim is mainly the development of multimodal model-based activity recogntion approaches. However, their visual sensor setup restricts them mostly to the domain of smart room. The CMU kitchen data set [12] contains videos of 43 actors preparing five different recipes, which were recorded from different camera views. Since the actors are wearing many invasive sensors during the recordings, their motions are influenced by the equipment which reduces the realism of the performed activities and also may disturb holistic activity recognition approaches. Because all of the five action sequence types can be actually tagged with one single activity *Cooking*, the data set has only a limited applicability for the evaluation of algorithms aiming at a household robot scenario, where many other activities are to be classified. In a similar fashion, the TUM Kitchen Data Set [13] also contains only one activity class, *Set Table*, which is performed by three subjects recorded in overhead views from four different angles. The POETICON and the OPPORTUNITY data sets consist of recordings of natural activities for which the actors were barely restricted in the way how to perform the activities. The focus of the rather small (6 activity classes, each performed three times) POETICON corpus lies on collaborative two person activities, while main feature of the OPPORTUNITY data set is its size, since it is to date the largest publicly available multimodal activity recognition data set. It consists of two types of recordings - an activities of daily living (ADL) run and a drill run - each performed by 12 different subjects. During one ADL run, nine long lasting activities are performed in a sequential way. The goal of the drill run is to generate many instances of simple actions and thus each subject repeats 20 times a sequence of such actions (eg., open a door, drink).


(a) countertop:fridge

(b) countertop:sink

(c) countertop:corner

(d) room:door

(e) room:window

Fig. 1. Sample images from the different camera views used in both setups.

As with the Robo-kitchen data set, the main application domain of the MINTA data set [16] are humanoid household robots. It contains recordings of six activitiy classes that are performed by each of the ten subjects ten times and also includes annotations of 60 temporally fine grained action primitives. However, its major drawback is, that it is set in a very simplified and thus unrealistic scenario. The University of Rochester Activities of Daily Living data set, published by Messing et al. [17], mostly resembles our Robo-kitchen data set. It consists of ten activity classes shot at a high resolution with a frame rate of 30 fps. Each activity is performed three times by each of the five different subjects in front of a non-uniform background and includes object interactions, as well as some variability with respect to the way the activities are performed and the subjects's appearance, i.e., age, gender, and ethnicity. Still, the performed activities are relatively short, simple and appear in some cases to be performed in an unnatural fashion. Also, in contrast to our work the data set contains neither stereo camera recordings nor multiple viewpoints.

Regarding action and activity recognition algorithms, researchers recently drew much motivation from approaches successfully applied to object detection and recognition problems. Especially, holistic approaches based on Space-Time Interest Points (STIP), which are a temporal extension of the popular local feature detectors, e.g., SIFT [18], gained a significant amount of attentionin the research community, e.g., [19]–[22]. In a similar fashion, many spatial local feature descriptors are extended to the temporal dimension, e.g., temporal versions of histograms of oriented gradients (HOG) [9], [23], SIFT [24], and SURF [21]. In order to capture motion information, many descriptors have been proposed based on optical flow, e.g., [8], [17], [25]. Usually, feature types encoding motion and appearance are used in a combined fashion, which is in fact, according to recent studies in neuro-science, in line with the way humans perceive motion patterns [26]. A comparative evaluation of different combinations of STIP detectors and descriptors can be found in [27] and [28]. The low-level features are either mapped directly to the appropriate action class, e.g., [8], [29], or first combined, e.g., with a bag-of-features model [30] and then classified with an SVM, as shown in, e.g., [9], [25].

Approaches dealing with the more complex task of classifying simple activities and person interactions can be found in e.g., [31], [32]. Wojek et al. investigated a multilayer HMM-based approach for the recognition of more complex office room activities, such as *Meeting* or *Paperwork* in [33]. Recently, Ryoo and Aggarwal [34] proposed a method to recognize multiple multi-person high-level activities by augmenting a bag-of-features model with spatial and temporal relations. In this paper, we also follow a bag-of-features paradigm as presented by Laptev et al. [9] but extend it for the recognition of complex activities. The extension is motivated by the work of Schindler et al. [35] in the field of action recognition, as well as our focus on recognizing kitchen activities consisting of a quasi-periodic repetition of atomic actions.

III. The KIT Robo-kitchen data set

In order to best capture the challenges that occur in the humanoid household robot domain, the Robo-kitchen data set[1] was specifically designed to match the setup of the robot ARMAR III [36]. As opposed to most current publicly available data for the evaluation of holistic human motion understanding approaches, the focus of the presented data set is to capture complex, long-lasting, quasi-periodic, and realistic kitchen activities. The recordings were conducted with multiple stereo cameras at a resolution of 640 × 480 pixels and a frame rate of 15 fps. The cameras were positioned at different locations in the room that are easily accessible by a robot platform. The use of multiple view-points allows for the evaluation of activity recognition approaches aiming at achieving robustness to view changes which we hope to achieve by exploiting depth from stereo. It is also expected that the depth information will improve activity recognition, since it allows to infer the 3D position of persons in the room, which is a strong prior on the likelihood of specific activities. Table I summarizes the technical details about the recorded data.

TABLE I
Overview of the Robo-kitchen data set.

Camera resolution	640 × 480 pixels
Frame rate	15 fps
Number of activities	14
Activities with 2 viewpoints	9
Activities with 3 viewpoints	5
Number of actors per activity	17

One of our main goals was that the activities were performed as natural as possible and thus, the subjects only got brief information about what to do, such as where to find the required objects, for how many people to set the table or to perform the activity at a location of their choice at the table. Each activity has been performed once by 17 different subjects of different age, gender, cultural background, and household skills in order to capture a hight amount of variation, as opposed to having only few actors repeating the activities several times. The duration of a video sequence varies between 10 s and 4 min, depending on the complexity of the activity and the thoroughness of the subject. For the recordings, two different camera setups have been used, one focussing on activities performed in the countertop area of the kitchen and the other for the whole room area. The reasoning for using two setups is that persons working at the countertop occlude the area where the activity takes place with their body when viewed from the room setup. In such cases, the robot should recognize that his position is not optimal for activity recognition and then move to a more suitable one.

Using the countertop setup, we recorded seven different activities, which are described including their canonical names in Tab. II. All of the activities have been recorded from three different view-points at the same time, with the exception of *Wash* and *Dry* because the camera in front of the sink had to be removed in order to allow access. It should be noted, that one of the cameras cannot be reached by a robot platform. However, since achieving robustness to view changes in activity recognition is an important, but still open topic, it has been added to the setup. Samples from the resulting views are given in Fig. 1 (a)-(c).

The recordings using the room setup are meant to model one of the primary applications of activity recognition for humanoid household robots. The key idea is that the robot takes the role of a servant observing the scene from a place where he has a good view over the room and his help if he assesses it might be required. Situation understanding is also important for the robot when entering a room in search for a new task to be performed. Note that only two camera views were used for the room recordings, but the positions of both are easily reachable by a robot platform. Figure 1 (d)-(e) contains examples of the field of view of the cameras used in this setup and Tab. III a list of the recorded activities. Many

[1] http://cvhci.anthropomatik.kit.edu/projects/act/kitchen

TABLE II
DESCRIPTION OF ACTIVITIES RECORDED USING THE "COUNTERTOP" SETUP.

Activity	Description	Seq. Length (s)	
		μ	σ
peel	Peeling some vegetables (carrots, cucumbers, potatoes) with a peeler.	137	66
cut	Slicing the peeled vegetables with a knife.	116	59
fry	Frying the sliced vegetables in a pan.	75	17
stir	Stirring of soup in a pot on the stove.	69	18
wipe	Wiping the countertop with a cloth.	34	24
wash	Washing the dishes in the sink.	133	64
dry	Drying and storing the washed dishes.	86	44

TABLE III
DESCRIPTION OF ACTIVITIES RECORDED USING THE "ROOM" SETUP.

Activity	Description	Seq. Length (s)	
		μ	σ
peel	Peeling some vegetables (carrots, potatoes) with a peeler.	118	70
cut	Slicing the peeled vegetables with a knife.	93	45
wipe	Wiping the table with a cloth.	90	19
set table	Setting the table for three people.	110	19
clear table	Moving dishes from the table to the dishwasher.	99	19
empty dishwasher	Emptying the dishwasher and storing the dishes and silverware in cupboards and drawers.	67	13
sweep	Sweeping the floor with a broom.	90	21
coffee	Reading a newspaper at the table and occasionaly sipping a cup of coffee.	149	47
pizza	Eating pizza with fork and knife.	70	61
soup	Eating soup with a spoon.	128	51

of the ten room activities involve walking around the whole kitchen area and perform different tasks at different locations. For example, the activity *Set Table* consists of opening/closing cupboards and drawers and several repetitions of picking up objects, transporting them to the table, and placing them at the proper place.

IV. ACTIVITY RECOGNITION

Motivated by the success of bag-of-features (BoF) representations of interest points for object classification, recently many researchers extended such approaches to the temporal domain with the outcome of state-of-the-art action recognition approaches. Following the same idea, our activity recognition system that is meant as a baseline for future research is based on a spatio-temporal extension of the Harris interest point detector [9]. Each such Space-Time Interest point (STIP) defines a three-dimensional volume within a video sequence whose extension depends on the size of the detected structure. To encode the local structure, we subdivide the volume in a grid of $(n_x = 3, n_y = 3, n_t = 2)$ cuboids, which are each described with a single normalized histogram of oriented gradients (HOG) and a histogram of optical flow (HOF). We calculate the HOG descriptors by applying a Sobel filter to the image data and then discretizing the gradient orientation to four equal sized classes. In a similar fashion, HOF features are calculated by binning the optical flow vectors within each sub-cuboid to a five-dimensional histogram with four bins representing motion direction and one bin when no motion is present. For the optical flow calculation, we use an implementation of the KLT tracker.

Before being mapped to an activity class, all low-level features calculated within a time interval are combined as a BoF for each feature type, i.e., HOG and HOF. We learn the codebook using k-means clustering with a random initialization of the cluster centers within the training data boundaries. To be less dependent on the initial cluster position, codebook learning is performed 30 times and the outcome of the training episode yielding the most compact clusters is taken. As suggested in [9], we subsample an equal amount of STIP descriptors from each training video in order to speed up the codebook building process. To build the BoF model, we calculate a histogram where each low-level feature contributes to the bin associated with the most similar (i.e., nearest) codebook word. The resulting HOG and HOF BoF representations are normalized and concatenated to form the feature vector for SVM classification. These vectors form the input for an SVM classifier using a χ^2-kernel and following a *one-vs-all* strategy to cope with multiple activity classes.

When used for action recognition, each video sequence is fully encoded with one BoF representation which is used for classification. However, this procedure is not feasible for a humanoid household robot, since it should obtain an estimate of the current situation online to ensure a very short response time. Motivated by the work of Kläser et al. [23] demonstrating that only few frames suffice to recognize simple actions, we base our system on short activity snippets. This is possible because we focus on recognising household activities that consist of a quasi-periodic repetition of simple action sequences. Thus, snippets of a length capturing at least one period should be meaningful enough for activity recognition. Since in the household robot scenario it is important to assess the current situation at any time, we use activity snippets located at each possible location within a video sequence for the experimental evaluation.

V. EXPERIMENTAL EVALUATION

For the experimental evaluation of the activity recognition system, we split the data in a development set, consisting of seven subject and used the sequences from the remaining ten subjects for testing purposes. Since the humanoid robot application of activity recognition requires that an estimation can be done at each time step, all possible activity snippets have been calculated as described in Sec. IV and used for classification. As proposed by Laptev et al. [9], we randomly sampled $100,000$ features to build the BoF codebooks and initialized the clustering algorithm with $4,000$ means. In order to speed up classifier training, we sampled 50 activity snippets from each training video. A set of suitable SVM kernel parameters has been determined with 10-fold cross-validation using the development data.

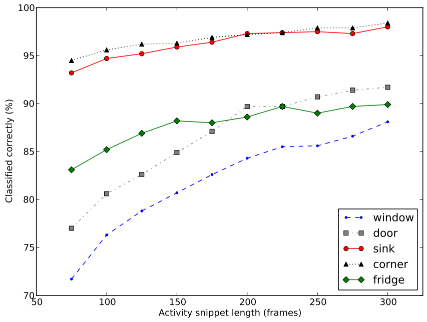
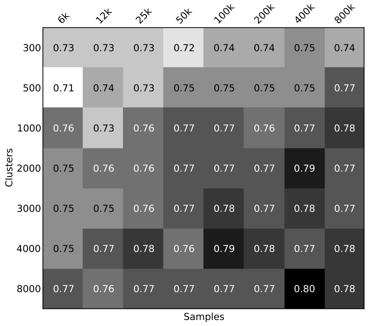


Fig. 2. Classification accuracy on the test set using different long activity snippets.

The most important question concerns a suitable length of the activity snippets, since the choice comes with a trade-off between classification accuracy and response-time for the robot. On the one hand, it is desired to get an estimate after a very short recording time, but on the other hand, longer snippets capture more information about the activity and thus lead to a higher classification rate. To investigate the impact of the snippet length on the classification accuracy, we evaluated the system starting with a length of 75 frames and increasing the length by 25 frames up to a total snippet size of 300 frames. As can be observed Fig. 2, in most views using more than 150 frames snippets does not improve the results enough to be worth the additional temporal overhead and thus we use this snippet length for further experiments. For future work, it is however important to lower the necessary snippet size while maintaining a high classification accuracy, in order to minimize the robot's response time.

Since, the codebook learning process also influences the system's performance, we further investigated the impact of the sample rate and codebook size on the classification accuracy. In Fig. 3 exemplary results from the room:door view with varying codebook parameters can be found. From this experiment, it can be inferred, that to a certain degree it is sufficient to either increase codebook size or the number of training samples, depending on the application. This is important when looking into achieving real-time activity recognition, since smaller codebooks also mean shortening feature calculation and classification time and thus improve runtime.

The final experimental results for each view, using 150 frame sized activity snippets and a 4000 words codebook can be seen in Tab. IV. It is not surprising, that the average classification accuracy for the room setup is lower than for the countertop setup, since it consists of more activity classes that are similar to each other in some cases. Exemplary confusion matrixes for one view each from the room and the countertop setup are to be found in Fig. 4 and Fig. 5 respectively. In both setups, the *cut* and *peel* activities are often confused, which is most likely due to the similarity of the motions they consist of. However, most of the remaining activities are recognized with an accuracy greater than 90%. It should be also noted, that the activities *cleartable* and *settable* are rarely confused, even though they mainly consist of the same actions, which are only performed in a slightly different order. All of these results support our hypothesis that the use of short activity snippets is sufficient for classification of complex quasi-periodic household activities.

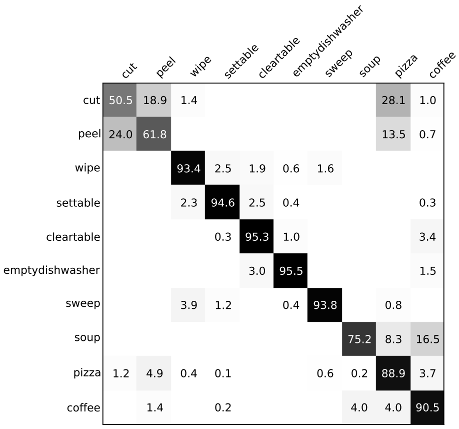


	6k	12k	25k	50k	100k	200k	400k	800k
300	0.73	0.73	0.73	0.72	0.74	0.74	0.75	0.74
500	0.71	0.74	0.73	0.75	0.75	0.75	0.75	0.77
1000	0.76	0.73	0.76	0.77	0.77	0.76	0.77	0.78
2000	0.75	0.76	0.76	0.77	0.77	0.77	0.79	0.77
3000	0.75	0.75	0.76	0.77	0.78	0.77	0.78	0.77
4000	0.75	0.77	0.78	0.76	0.79	0.78	0.77	0.78
8000	0.77	0.76	0.77	0.77	0.77	0.77	0.80	0.78

Fig. 3. Resulting accuracies for different combinations of sample and codebook size (10 times run average, room:door view, 150 frames long snippets).

TABLE IV
CLASSIFICATION ACCURACIES ON THE TEST SET FOR EACH SETUP USING 150 FRAMES LONG SNIPPETS.

setup	countertop			room	
camera	corner	fridge	sink	door	window
accuracy	96.3 %	88.2 %	95.9 %	84.9 %	80.7 %
activities	7	7	5	10	10



	cut	peel	wipe	settable	cleartable	emptydishwasher	sweep	soup	pizza	coffee
cut	50.5	18.9	1.4						28.1	1.0
peel	24.0	61.8							13.5	0.7
wipe			93.4	2.5	1.9	0.6	1.6			
settable			2.3	94.6	2.5	0.4				0.3
cleartable				0.3	95.3	1.0				3.4
emptydishwasher					3.0	95.5				1.5
sweep			3.9	1.2		0.4	93.8		0.8	
soup								75.2	8.3	16.5
pizza	1.2	4.9	0.4	0.1			0.6	0.2	88.9	3.7
coffee		1.4		0.2				4.0	4.0	90.5

Fig. 4. Confusion matrix for the room:door view (test set, 150 frames long snippets).

VI. CONCLUSION AND FUTURE WORK

This paper has introduced a novel data set consisting of realistic, complex kitchen activities recorded in a setting closely resembling a humanoid household robot scenario. We hope that the presented Robo-kitchen data set will provide a basis for the development and comparison of activity recognition approaches aiming at applications in the household robotics

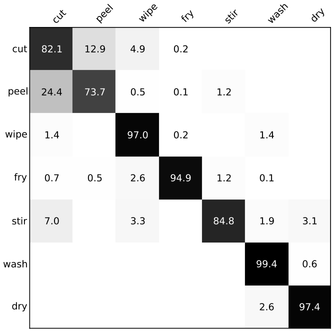


Fig. 5. Confusion matrix for the countertop:fridge setup (test set, 150 frames long snippets).

domain. And we believe this topic to be an important component in enhancing a natural HRI, which however has not attracted much attention in the research community yet.

We further extended a state-of-the-art action recognition approach for activity classification based on the idea that typical kitchen activities consist of a quasi-periodic repetition of atomic actions. The system is evaluated using the presented Robo-kitchen data set with the focus on future developments, such as lowering the response time of the robot and achieving a real-time capability for the recognition system.

In our future work, we intend to address the problems of robustness against view-point changes. We will also investigate how depth from stereo can be exploited to improve the activity recognition accuracy. Since stereo-cameras have been used for the recordings, it is a logical step to use the depth information in order to enhance activity recognition by context estimation through 3D person tracking.

Acknowledgment

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center SFB 588 on “*Humanoid Robots – Learning and Cooperating Multimodal Robots*”and by the German Federal Ministry of Education and Research (BMBF) as part of the GEMS program.

References

[1] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *CVIU*, 2006.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, “Machine recognition of human activities: A survey,” *IEEE Transactions on Circuits and Systems*, 2008.

[3] R. Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing*, 2010.

[4] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *ICPR*, 2004.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *ICCV*, 2005.

[6] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *CVIU*, 2006.

[7] L. Sigal and M. J. Black, “HumanEva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion,” Tech. Rep., 2006.

[8] I. Laptev and P. Pérez, “Retrieving actions in movies,” *ICCV*, 2007.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *CVPR*, 2008.

[10] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in *CVPR*, 2009.

[11] J. Liu, L. Jiebu, and M. Shah, “Recognizing realistic actions from videos in the wild,” in *CVPR*, 2009.

[12] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, and J. Macey, “Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database,” Tech. Rep., 2009.

[13] M. Tenorth, J. Bandouch, and M. Beetz, “The TUM Kitchen Data Set of everyday manipulation activities for motion tracking and action recognition,” in *ICCV*, 2009.

[14] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. Millan, “Collecting complex activity datasets in highly rich networked sensor environments,” in *INSS*, 2010.

[15] C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, and K. Pastra, “The POETICON enacted scenario corpus - a tool for human and computational experiments on action understanding,” in *FG*, 2011.

[16] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen, “Combined intention, activity, and motion recognition for a humanoid household robot,” in *IROS*, 2011.

[17] R. Messing, C. Pal, and H. Kautz, “Activity recognition using the velocity histories of tracked keypoints,” *ICCV*, 2009.

[18] D. G. Lowe, “Object recognition from local scale-invariant features,” in *ICCV*, 1999.

[19] I. Laptev and T. Lindeberg, “Space-time interest points,” in *ICCV*, 2005.

[20] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior Recognition via Sparse Spatio-Temporal Features,” in *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2006.

[21] G. Willems, T. Tuytelaars, and L. V. Gool, “An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector,” in *ECCV*, 2008.

[22] T.-H. Yu, T.-K. Kim, and R. Cipolla, “Real-time Action Recognition by Spatiotemporal Semantic and Structural Forests,” in *British Machine Vision Conference*, 2010.

[23] A. Kläser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3D-gradients,” in *BMVC*, 2008.

[24] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” *ACM Multimedia*, 2007.

[25] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, “Action Recognition by Dense Trajectories,” in *CVPR*, 2011.

[26] M. A. Giese and T. Poggio, “Neural mechanisms for the recognition of biological movements.” *Nature reviews. Neuroscience*, 2003.

[27] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *British Machine Vision Conference*, 2009.

[28] J. Stöttinger, B. T. Goras, T. Pöntiz, A. Hanbury, N. Sebe, and T. Gevers, “Systematic Evaluation of Spatio-temporal Features on Comparative Video Challenges,” in *ACCV Workshop on Video Event Categorization, Tagging and Retrieval*, 2010.

[29] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz, “HMM-based human motion recognition with optical flow data,” in *Humanoids*, 2009.

[30] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[31] S. Hongeng, R. Nevatia, and F. Bremond, “Video-based event recognition: activity representation and probabilistic recognition methods,” *CVIU*, 2004.

[32] M. S. Ryoo and J. K. Aggarwal, “Recognition of Composite Human Activities through Context-Free Grammar based Representation,” in *CVPR*, 2006.

[33] C. Wojek, K. Nickel, and R. Stiefelhagen, “Activity Recognition and Room-Level Tracking in an Office Environment,” in *Intl. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[34] M. S. Ryoo and J. K. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” *ICCV*, 2009.

[35] K. Schindler and L. Van Gool, “Action Snippets: How many frames does human action recognition require?” in *CVPR*, 2008.

[36] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, “ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control,” in *Humanoids*, 2006.