Automatic task assistance for people with cognitive disabilities in brushing teeth - a user study with the TEBRA system

CHRISTIAN PETERS, THOMAS HERMANN, SVEN WACHSMUTH, Cognitive Interaction Technology - Center of Excellence, Bielefeld University, Germany
and JESSE HOEY, David Cheriton School of Computer Science, University of Waterloo, Canada
, Toronto Rehabilitation Institute, Toronto, Canada

People with cognitive disabilities such as dementia and intellectual disabilities tend to have problems in coordinating steps in the execution of Activities of Daily Living (ADLs) due to limited capabilities in cognitive functioning. To successfully perform ADLs, these people are reliant on the assistance of human caregivers. This leads to a decrease of independence for care recipients and imposes a high burden on caregivers. Assistive Technology for Cognition (ATC) aims to compensate for decreased cognitive functions. ATC systems provide automatic assistance in task execution by delivering appropriate prompts which enable the user to perform ADLs without any assistance of a human caregiver. This leads to an increase of the user's independence and to a relief of caregiver's burden. In this article, we describe the design, development and evaluation of a novel ATC system. The TEBRA (TEeth BRushing Assistance) system supports people with moderate cognitive disabilities in the execution of brushing teeth. A main requirement for the acceptance of ATC systems is context awareness: explicit feedback from the user is not necessary to provide appropriate assistance. Furthermore, an ATC system needs to handle spatial and temporal variance in the execution of behaviors such as different movement characteristics and different velocities. The TEBRA system handles spatial variance in a behavior recognition component based on a Bayesian network classifier. A dynamic timing model deals with temporal variance by adapting to different velocities of users during a trial. We evaluate a fully functioning prototype of the TEBRA system in a study with people with cognitive disabilities. The main aim of the study is to analyze the technical performance of the system and the user's behavior in the interaction with the system with regard to the main hypothesis: is the TEBRA system able to increase the user's independence in the execution of brushing teeth?

CCS Concepts: •**Human-centered computing** → **Empirical studies in HCI;** •**Social and professional topics** → **People with disabilities;**

General Terms: Experimentation, Human factors

Additional Key Words and Phrases: Automatic task assistance, Cognitive disabilities, User study

ACM Reference Format:
Christian Peters, Thomas Hermann, Sven Wachsmuth and Jesse Hoey, 2013. Automatic task assistance for people with cognitive disabilities in brushing teeth - a user study with the TEBRA system *ACM Trans. Access. Comput.* V, N, Article A (January YYYY), 35 pages.
DOI: http://dx.doi.org/10.1145/0000001.0000001

1. INTRODUCTION

People with cognitive disabilities form a primary group of healthcare recipients due to their limited capabilities in cognitive functioning such as perception, reasoning and remembering [Gillespie et al. 2011]. Problems related to this functioning appear in a human's daily routine where the successful execution of Activities of Daily Living

This work has been supported by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.

DOI: http://dx.doi.org/10.1145/0000001.0000001

(ADLs) is an integral part of an autonomous and self-determined life. ADLs refer to everyday tasks such as eating, dressing or personal hygiene. One problem for people with cognitive disabilities in the execution of ADLs is task sequencing. Task sequencing refers to the ability to decompose tasks into sub steps. For a successful execution of the overall task, the sub steps need to be combined in an appropriate order. For most tasks such as hand washing, tooth brushing and dressing, the sub steps can be combined in a flexible way which allows for different ways of task execution. In a dressing task, for example, a user might put on the left sock first and the right sock afterwards, or vice versa.

Flexibility in task execution imposes a high risk of erroneous behavior for people with cognitive disabilities: users forget steps or get stuck in task execution. In such cases, an external intervention of a human caregiver is necessary for a proper task execution. Hence, an inability to perform ADLs leads to a decrease or even a loss of independence and makes people with cognitive disabilities highly dependent on a human caregiver. Furthermore, an inability to perform ADLs might impose security risks for the well-being of people: for example, a person with Alzheimer's disease aims to prepare tea, but forgets to pour water into the kettle. Turning on a kettle without any water inside is a potential fire hazard. Professional caregivers as well as informal caregivers such as family members worry about the well-being of the care recipients. This leads to an emotional burden for the caregiver which might result in chronic stress and consequent diseases [Bevans and Sternberg 2012].

Assistive Technology for Cognition (ATC) refers to technical interventions which compensate for decreased or missing cognitive capabilities by providing prompts which assist the user in the execution of ADLs. The application of ATC aims at increasing the independence of people with cognitive disabilities from a human caregiver. This leads to an increase in self-esteem and self-determination in a care recipient's life and, furthermore, to a relief of caregiver burden due to the prolonged independence of the care recipient [Scherer et al. 2005].

The main goal of ATC systems is to foster the independence of the user by providing appropriate prompts when necessary for a successful task execution. A prompt is necessary in three situations: firstly, a person might forget a step in the overall task which leads to inappropriate follow-up behaviors. For example, a user rinses hands in a hand washing task without having taken soap first. Secondly, a person might not be able to terminate a sub step of the task due to obsessive behavior. Thirdly, a person is not able to focus on the task and loses track of the overall progress due to environmental distractions such as noise. In these situations, a prompt is necessary to assist the user in task execution.

Context awareness enables a system to detect such situations without explicit feedback of the user about completed steps: a context aware system infers a user's current behavior as well as the overall progress in the task based on sensory information obtained in the environment. The implementation of a context-aware behavior is difficult since an ATC system needs to deal with the spatial and temporal variance that people with cognitive disabilities show in task execution. In this article, we refer to spatial and temporal variance in the context of a user's behavior in ATC: spatial variance refers to differences in the execution of behaviors due to different motor abilities which result in different movement characteristics amongst individual users. Temporal variance denotes differences in the velocities of task execution which may vary greatly between individual users. For example, one user might perform sub steps of a task very slow, but another user might be very quick in execution.

In this article, we describe the design, development and evaluation of a novel context-aware ATC system which is robust with regard to spatial and temporal variance of users: the *TEBRA* system (TEeth BRushing Assistance system) assists people

with cognitive disabilities in the execution of *brushing teeth*. Brushing teeth is an important basic ADL since (1) disregarding oral hygiene can lead to severe medical problems and (2) people with cognitive disabilities usually have problems with brushing teeth due to the flexibility and complexity of the task: brushing teeth involves several objects such as paste and brush which are used in different sub steps during the task. The sub steps can be combined in a flexible way for successful task execution.

To handle the spatial variance of people with cognitive disabilities we infer a user's behaviors based on the states of objects manipulated during the behaviors. The behavior recognition component is described in section 4.1. We deal with the temporal variance in task execution by using a dynamic timing model with a number of different thresholds that are automatically adjusted during a trial. Section 4.2.2 describes the timing model in detail.

The target group of users are people with moderate cognitive disabilities such as intellectual disabilities and Autistic Spectrum Disorder, but also age-related disabilities such as dementia and Alzheimer's disease. We cooperate with the residential home *Haus Bersaba* belonging to *v. Bodelschwinghsche Stiftungen Bethel*, a clerical foundation in Bielefeld, Germany. 35 people with mild to moderate cognitive disabilities live in *Haus Bersaba* and receive permanent care by professional caregivers including assistance in brushing teeth: a caregiver stands beside the person and assists during the brushing task by providing verbal and visual prompts.

In a study with target group users, we evaluate the technical system performance including the recognition and tracking of a user's behaviors in the overall task as well as the appropriateness of prompts. Furthermore, we analyze the user's reactions to prompts and discuss aspects of usability and acceptance of the TEBRA system.

We do not address task initiation (getting a person to *start* the task) in this study. Although this is an important aspect, our study follows a similar protocol as that in [Mihailidis et al. 2008], and we leave task intitiation for future work.

The article is structured as follows: Section 2 discusses related work. Section 3 describes the design of the system. A technical overview of the TEBRA system is given in section 4, which may be skipped by readers more interested in the user study aspects of the work. Section 5 describes the design of user study and discusses the results. We conclude the article in section 6.

2. RELATED WORK

In recent years, there were several attempts to classify ATC interventions according to cognitive functions for which the ATC compensates for [LoPresti et al. 2004; Gillespie et al. 2011]. LoPresti et al. distinguish between ATC compensating for executive function impairments and information processing impairments [LoPresti et al. 2004]. Gillespie et al. provide a systematic classification according to the International Classification of Functioning, Disability and Health (ICF) [Gillespie et al. 2011]. Gillespie et al. identified applications of ATC in the following areas of cognitive functions referring to the ICF classification: attention, calculation, emotional functions, experience of self and time, and higher-level cognitive functions.

This article describes a contribution by designing, developing and evaluating the TEBRA system which compensates for missing or decreased higher-level cognitive functions.

Most ATC identified by Gillespie et al. assist in one of the two areas of higher-level cognitive functions: organization and planning, and time management. Time management refers to scheduling a user's daily routine dealing with temporal constraints between different tasks. For example, PEAT (Planning and Execution Assistant Trainer) is a scheduling aid for people with brain injury [Levinson 1997]. PEAT structures a user's daily routine by providing visual and audible cues using a mobile phone. Simi-

lar to the PEAT system, Autominder schedules daily activities of people with mild to moderate memory impairments [Pollack et al. 2003]. Autominder models a user's daily plan, tracks a user's execution of the plan and decides whether to provide a prompt to the user. Both PEAT and Autominder are able to recognize conflicting tasks and to replan when the daily plan is modified.

In comparison to time management, organization and planning relates to the execution of single tasks in which the subtasks need to be structured and performed in a temporal order for a successful task execution. The GUIDE system assists people with cognitive disabilities in task execution [O'Neill and Gillespie 2008] by simulating the verbal assistance provided by a human caregiver. The system is able to understand simple verbal responses such as 'yes', 'no' or 'done' and provides assistance according to the user's responses. O'Neill et al. conducted a study with eight users who were assisted in donning a prosthetic limb [O'Neill et al. 2010]. Six of eight users showed a significant increase in task performance with a decreased number of errors and omissions during system assistance.

The COACH system assists people with mild to moderate dementia in the task of handwashing [Hoey et al. 2010]. COACH uses computer vision techniques for environmental perception and a Partially Observable Markov Decision Process (POMDP) for planning and decision making. The COACH system was evaluated in a user study with 6 participants having moderate-to-severe dementia [Mihailidis et al. 2008]. The participants' performance was tested in two alternating conditions: (1) baseline without COACH system, and (2) intervention with COACH system. The average rate of hand washing steps completed independently was increased by 11% in the inverventionn compared to the baseline scenario. Furthermore, intervention of a caregiver was decreased by 60% when using the COACH system. The COACH system was also applied to the task of prompting a person with a cognitive disability through a simple factory assembly process [Melonis et al. 2012].

In comparison to donning a limb and washing hands, the task of brushing teeth is more complex and flexible: according to the results of a task analysis technique described in section 3.1, brushing teeth consists of eight sub steps (*paste_on_brush*, *rinse_mug_fill*, *rinse_mug_clean*, *rinse_mouth_clean*, *rinse_mouth_wet*, *brush_teeth*, *clean_brush* and *use_towel*). A successful execution of the task involves the manipulation of four objects (brush, paste, mug and towel). In comparison, washing hands consists of five sub steps (wet hands, take soap, water on, water off, dry hands) in which two objects (soap, towel) are involved.

In the COACH system, a user's behavior is recognized implicitly using pairs of pre/post-actions: for example, if the hands enter (pre action) and leave (post action) the soap region, COACH infers that the user has taken the soap with a pre-specified probability. In the TEBRA system, we explicitly recognize the different behaviors involved in the brushing task and model the timing dynamics of behaviors in a timing model. Hence, the TEBRA system is more robust with regard to the spatial and temporal variance in task execution compared to other ATC systems.

Furthermore, the target group in the study described in this article is heterogeneous since people with different cognitive abilities such as Autistic Spectrum Disorder and intellectual disabilities participate. Hence, the study with the TEBRA system describes a contribution in the field of ATC by assisting a more heterogeneous user group in a more complex and flexible task than previous ATC systems.

3. SYSTEM DESIGN

The demands and abilities of users play an important role for the acceptance of ATC in a user's everyday life. According to Scherer et al., psychosocial factors such as disregarding the user's requirements during the design process is a common reason why

ATC application is abandoned quickly after deployment [Scherer et al. 2005]. User-centered design is a methodology[1] which incorporates the users' demands and abilities early into the design process [Gould and Lewis 1985].

Design decisions need to take into account the characteristics of the task. Task analysis is an important set of techniques to reveal task characteristics and provide initial design decisions. In the design of the TEBRA system, we use a task analysis method called Interaction Unit (IU) analysis which reveals characteristics of the brushing task using a structured methodology as described in the following section.

3.1. Interaction Unit analysis

Designing an ATC system based on common-sense knowledge about brushing teeth is not sufficient. Users of ATC are people with cognitive disabilities who usually show special characteristics in task execution: firstly, due to decreased motor abilities which often coincide with cognitive disabilities [Kluger et al. 1997], target group users might show uncommon usage of objects. We aim to take into account such differences to common behavior as far as possible in the design of the TEBRA system. Secondly, people living in a residential home commonly rely on the assistance of a caregiver while brushing their teeth. Caregivers aim to impart a routine in the execution of the brushing task which suits the user's abilities. We analyze the caregiver's way of task assistance and consider important aspects in the design phase. We conduct a qualitative data analysis on in-situ observations made at the residential home *Haus Bersaba* where people with moderate cognitive disabilities permanently live. In-situ observations are a common way to study a user's behavior in a natural environment [Leroy 2011; Intille et al. 2004]. Each observation is a video which shows a user brushing teeth while being observed and supported by a caregiver. Figure 1 depicts an example image. We recorded 23 trials performed by eight users at three different days where seven

Fig. 1. Example image of the in-situ observation.

users conducted three trials each and one user conducted two trials [Peters et al. 2011]. The users are supported by two caregivers assisting in 10 and 13 trials, respectively. We use Interaction Unit (IU) analysis proposed by Ryu and Monk as a method of task analysis [Ryu and Monk 2009]. IU analysis models user-machine interaction with cycles of interaction called interaction units. A user executes actions in order to achieve a desired goal. Actions are triggered using both visible cues of the environment and mental processes of a user. IU analysis describes actions, goals, environmental states

[1] according to ISO standard *Human-centered design for interactive systems* (ISO 9241-210, 2010)

and mental process in a single model and allows for a description of "the intimate connection between goal, action, and the environment in user-machine interaction" [Ryu and Monk 2009, p. 1]. Hoey et al. use an adapted form of IU analysis to facilitate the specification process of an automatic prompting system using a POMDP [Hoey et al. 2011]. We use a similar form of IU analysis to extract task-relevant information which we incorporate in the design of the TEBRA system. The results of IU analysis, given in table I, were obtained by iteratively analyzing the recorded videos. We decompose the brushing task into seven sub tasks given in column *UB*. We will refer to the sub tasks as *user behaviors* in the following. User behaviors are *paste_on_brush*, *fill_mug*, *rinse_mouth*, *brush_teeth*, *clean_mug*, *clean_brush* and *use_towel*. Column *Current goals* describes a user's goal stack where *Final* means the overall goal of getting the teeth brushed properly. Whenever a user behavior is initiated, the behavior is added to the goal stack as the user's current goal. When the user behavior is completed, the goal is removed from the stack and *Final* is the current goal again. Each user behavior is further subdivided into single steps described in column *UB steps*. For example, performing *rinse_mouth* consists of a sequence of three steps: mug is moved to the face, the user rinses his/her mouth and the user moves the mug away from the face. Column *Mental processes* describes the mental processes involved to initiate user behavior steps (these are called *abilities* in [Hoey et al. 2011]). Ryu and Monk distinguish between three mental processes: *recognition*, *recall* and *affordance*.

Recognition (Rn). Recognition means that the user can directly perceive an object's state in the environment, e.g. mug is empty in IU 2 in table I.
Recall (Rl). The user needs to remember a certain state of the environment which is not directly observable. For example, the user has to recall that the mug is dirty in IU 18 because it was used in a previous step.
Affordance (Af). Affordance describes the recognition of the meaning of an object and the way to use it, e.g. the tap can be altered to on which makes the water flow in IU 20.

Column *Current environment* describes the environmental configuration as preconditions of single user behavior steps. Performing the step changes the environmental configuration, for example in the first step of *paste_on_brush*: the toothpaste is on the counter and taking the paste changes the location to 'in hand'. We utilize the environmental configuration given in column *Current environment* to extract environmental states that we encode in discrete variables as depicted in table II. We distinguish between *behavior* and *progress* variables: we apply *behavior* variables to recognize user behaviors in a recognition component. The *progress* variables are hard to observe using sensory information due to reasons of robustness: for example, it is very error-prone to visually detect whether the *brush_condition* is *dirty* or *clean*. A specialized sensor at the brushing head is not desirable due to hygienic reasons. However, the *progress* variables are important since they are part of the environmental state during the task. We utilize *progress* variables to monitor the user's progress in brushing teeth which is described later in this section. We abstract from the recognition of single behavior steps as given in column *UB steps* in table I. Instead, we infer the user's behavior based on the *behavior* variables which express states of objects manipulated during a behavior. From column *Current environment*, we extract five *behavior* variables describing important object states: *mug_position*, *towel_position*, *paste_movement*, *brush_movement* and *tap_condition*. The upper part of table II shows the five variables and their according discrete values. For *brush_movement*, we have the values *no*, *yes_sink* and *yes_face*. The latter ones are important to discriminate between the user behaviors *paste_on_brush* and *brush_teeth* based on the movement of the brush. The values of the variables *mug_position* and *towel_position* are the different regions identified in

Table I. Results of the IU analysis for brushing teeth. TT = toothpaste tube, Rn = Recognition, Rl = Recall, Af = Affordance. See text for a detailed description of the table.

UB	IU	Current goals	Current environment	Mental processes	UB steps
fill_mug	1	Final	mug on counter	Rn mug on counter Rl step	no action
	2	Final, fill_mug	mug empty	Rn mug empty Af tap	give mug to tap
	3	Final, fill_mug	mug at tap, tap off	Af tap on	alter tap to on
	4	Final	mug at tap, tap on	Af tap off	alter tap to off
		Final	mug filled		
rinse_mouth	5	Final	mug filled	Rl step	no action
	6	Final, rinse_mouth	mug filled	Af mug	give mug to face
	7	Final, rinse_mouth	mug at face	Af mug	give water to mouth
	8	Final	mug else	Af counter	give mug to counter
		Final	mug counter		
paste_on_brush	9	Final	brush on counter TT on counter TT on counter	Rn brush Rn TT on counter Rl step	no action
	10	Final, paste_on_brush	TT on counter	Af TT	take TT from counter
	11	Final, paste_on_brush	brush on counter	Af brush	take brush from counter
	12	Final, paste_on_brush	brush and TT in hand	Af TT	spread paste on brush
	13	Final	TT in hand	Af counter	give TT to counter
		Final	TT on counter, brush in hand		
brush_teeth	14	Final	brush with paste in hand	Af brush Rl step	no action
	15	Final, brush_teeth	brush with paste in hand	Af brush	give brush to face
	16	Final, brush_teeth	brush at face	Af brush	brush all teeth
	17	Final	brush at face, teeth clean	Rl teeth clean	take brush from face
		Final	brush not at face		
clean_mug	18	Final	mug dirty at counter	Rl mug dirty Rl step	no action
	19	Final, clean_mug	mug dirty at counter	Rn mug dirty, Af tap	give mug to tap
	20	Final, clean_mug	mug dirty at tap, tap off	Af tap on	alter tap to on
	21	Final, clean_mug	mug dirty at tap, tap on	Rn water on, Af tap	give mug to tap
	22	Final	mug clean at tap, tap on	Af tap off	alter tap to off
	23	Final	mug clean at tap, tap off	Af counter	give mug to counter
		Final	mug clean at counter		
clean_brush	24	Final	brush dirty	Rn brush dirty Rl step	no action
	25	Final, clean_brush	brush dirty	Rl brush dirty	give brush to tap
	26	Final, clean_brush	brush dirty at tap, tap off	Af tap on	alter tap to on
	27	Final, clean_brush	brush dirty at tap, tap on	Rn water on, Af tap	give brush to tap
	28	Final	brush clean at tap, tap on	Rn water on, Af tap off	alter tap to off
	29	Final	brush clean at tap, tap off	Af counter	give brush to counter
		Final	brush clean at counter		
use_towel	30	Final	towel at hook, mouth wet	Rn mouth wet Rl step	no action
	31	Final, use_towel	towel at hook, mouth wet	Af towel	give towel to face
	32	Final, use_towel	towel at face, mouth wet	Af towel	dry mouth
	33	Final	towel at face, mouth dry	Af hook	give towel to hook
		Final	towel at hook		

Table II. *Behavior* and *progress* variables extracted from the environmental configuration in table I.

State variable	Values
behavior	
mug_position	counter, tap, face, else, no_hyp
towel_position	hook, face, else, no_hyp
paste_movement	no, yes
brush_movement	no, yes_sink, yes_face
tap_condition	off, on
progress	
mug_content	empty, water
mug_condition	dirty, clean
mouth_condition	dry, wet, foam
brush_content	no_paste, paste
brush_condition	dirty, clean
teeth_condition	dirty, clean

column *Current environment* where the mug and towel appear during task execution. *No_hyp* is used if no hypothesis about the mug/towel position is available. The lower part of table II shows *progress* variables and their according discrete values which we use to monitor the user's progress in the task. At each time in task execution, the user's progress is modeled by the set of six *progress* variables which we will denote *progress state space* in the following. The occurrence of a user behavior during the execution of the task leads to an update of the progress state space: we define necessary preconditions and effects of user behaviors in terms of *progress* variables. When a user behavior occurs, we check whether the preconditions are met and, if so, update the progress state space with the effects of the current behavior. Table III shows the preconditions and effects for user behaviors in terms of *progress* variables extracted during IU analysis. We distinguish between *rinse_mouth_wet* and *rinse_mouth_clean*: the behaviors

Table III. Preconditions and effects of user behaviors extracted from the environmental configuration in table I.

User behavior	Preconditions	Effects
paste_on_brush	brush_content=no_paste teeth_condition=dirty	brush_content=paste brush_condition=dirty
fill_mug	mug_content=empty	mug_content=water
clean_mug	mug_content=empty mug_condition=dirty teeth_condition=clean	mug_condition=clean
rinse_mouth_clean	mug_content=water mouth_condition=foam teeth_condition=clean	mug_condition=dirty mouth_condition=wet mug_content=empty
rinse_mouth_wet	mug_content=water mouth_condition=dry	mug_condition=dirty mouth_condition=wet
brush_teeth	brush_content=paste teeth_condition=dirty mouth_condition=wet	teeth_condition=clean brush_content=no_paste mouth_condition=foam brush_condition=dirty
clean_brush	brush_condition=dirty teeth_condition=clean	brush_condition=clean brush_content=no_paste
use_towel	mouth_condition=wet teeth_condition=clean	mouth_condition=dry

Fig. 2. Washstand setup equipped with sensor technology.

are equal with regard to object usage, but differ in the semantics based on the time at which the behaviors are executed within the overall task. Video analysis showed that wetting the mouth with water using the mug (before brushing the teeth) is a common step as part of the user's regular daily routine. If a user forgets this steps, the caregiver will intervene and prompt the user to do so. This step is described as *rinse_mouth_wet* whereas cleaning the mouth after the brushing step is *rinse_mouth_clean*. The preconditions and effects of *rinse_mouth_wet* and *rinse_mouth_clean* differ. Hence, we differentiate between these behaviors in tracking a user's overall progress in the task. The main findings of the IU analysis are three-fold: firstly, we decomposed the brushing task into eight user behaviors given in table III. Secondly, we identified variables as given in table II which describe important objects and according discrete states that are relevant during task execution. Thirdly, we determined preconditions and effects of user behaviors shown in table III in order to track a user's progress in the task. In the following section, we describe the construction of the washstand setup and the equipment of the setup with sensor technology in order to recognize behaviors identified in the IU analysis.

3.2. Setup and sensors

We built a washstand setup as depicted in figure 2. We installed a customary washbowl with a single-lever mixer tap and a mirror. All installations comply with the DIN 18024-2 norm for sanitary areas which are accessible for people with impairments. We

equipped the washstand with a TFT display including speakers as a device to prompt the user during task execution. As shown in figure 2, the TFT display is installed between the mirror and the sink. We integrated the prompting device into the setup in a central position because we don't want to shift the user's attention away from the washstand during prompting. In order to recognize the user behaviors identified in the IU analysis as given in table I, the washstand is equipped with a set of unobtrusive sensors for environmental perception. Unobtrusive means that the sensors are smoothly integrated into the environment without attaching sensors to the user's body directly. We avoid such wearable sensors because we don't want to disturb the user in the execution of the task.

The equipment of the washstand setup with environmental sensors is sensitive with regard to privacy concerns. Privacy issues arise due to the retrieval and storage of sensitive personal data in a user's bathroom. In the design and development process of the TEBRA system, storing a user's data is necessary to evaluate and enhance system performance. We obtained the user's declaration of consent before collecting sensitive data throughout the studies described in this article.

We equipped the washstand with two cameras to visually capture the important areas involved in tooth brushing: one camera observes the environment from an overhead perspective and captures the counter and the sink region. A second camera with a frontal perspective observes the upper body part of the user including the face. Figure 3 shows example images. According to table II, the state of the tap (*tap_condition*) and

Fig. 3. View perspectives of the frontal camera (left image) and the overhead camera (right image).

the toothbrush (*brush_movement*) are important for the recognition of user behaviors in tooth brushing. In order to determine the *tap_condition*, we installed a flow sensor (Gentech FCS-03) at the water supply to the tap. The flow sensor measures the water flow and provides a binary on/off signal. In order to distinguish between the three states of the *brush_movement* variable, we installed a sensor module into a commercially available, electric toothbrush as shown in figure 4. The brush is equipped with an x-imu sensor module manufactured by x-io technologies[2] as shown in the bottom right of figure 4. The sensor module has nine degrees of freedom: a gyroscope measuring the angular velocity of the change in orientation, an accelerometer providing gravitational acceleration and a magnetometer measuring the earth's magnetic field

[2]http://www.x-io.co.uk/

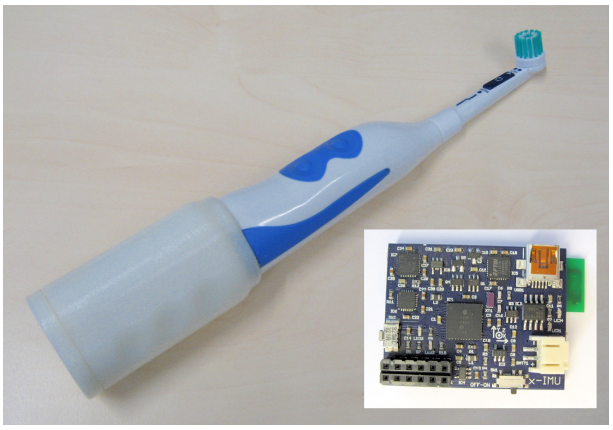


Fig. 4. Electric toothbrush used in the TEBRA system. The x-imu sensor shown in the bottom right of the image is installed into the handle of the brush.

in x,y and z-axis each. The x-imu unit is integrated into the handle of the brush[3] and provides a Bluetooth connection for wireless data transfer.

3.3. Interviews with caregivers

A successful prompt in an ATC system needs to be suitable in modality and level of information in a way that the user can understand and react correctly to the prompt [Seelye et al. 2012]. We aimed to find out about appropriate modalities and levels of information of prompts by conducting interviews with three caregivers of *Haus Bersaba*. Caregivers are experts in prompting since they provide professional nursing care in the task of brushing teeth as part of their daily routine. We interviewed the caregivers independently of each other and recorded the interviews in order to evaluate the caregiver's answers. In each interview, we presented prompts of three modalities: audio prompts, visual prompts and audio-visual combinations.

Audio. We chose an audio modality due to two reasons: firstly, users are familiar with audio prompts since caregivers mainly use verbal instructions. Secondly, ONeill and Gillespie argue that "prompting in the verbal medium rather than the visual medium provides a more direct augmentation of executive function" due to a close relationship between language and executive function in the human brain [O'Neill and Gillespie 2008, p. 9]. We used audio prompts in terms of verbal commands which were prerecorded by the first author of the article. We presented commands with different levels of detail ranging from short, specific commands (e.g. *"Clean mug."*) to more sophisticated instructions (e.g. *"Please, clean the mug in front of you."*). We asked the caregivers about different properties of the commands: (1) Is a male or a female voice more appropriate for prompting? (2) Is an unknown or a known voice more suitable?

Visual. Visual prompts are cognitively more demanding than audio prompts since they might shift the user's attention away from the task. However, visual prompts can be very effective since a wide range of visualizations from simple cues such as images to dense information presentations such as videos are possible. We presented two types of task-related visualizations to the caregivers including different levels of information: images of objects aim to trigger the user's memory and activate a user's routine of task

[3]This work was done by Simon Schulz from the Central Lab Facilities (CLF) of the Cognitive Interaction Technology Center of Excellence (CITEC) at Bielefeld University.

execution by giving appropriate hints. We presented pictograms showing a behavior, cartoon-like images and images of real-life objects. A video comprises much more information in a single prompt than an image: we recorded videos which show the first author of the article performing a behavior. Hence, the user can directly follow the behavior shown in the video which constitutes a more direct way of assistance. Figure 5 depicts a selection of visual prompts which were presented to the caregivers in the interviews.

Audio-visual. Audio-visual prompts are combinations of the above-mentioned audio and visual prompts, e.g. a cartoon-like image paired with a verbal command. As a special type of audio-visual prompts, we augmented an audio command with embodiments of prompts such as a virtual agent or a cartoon-like character.

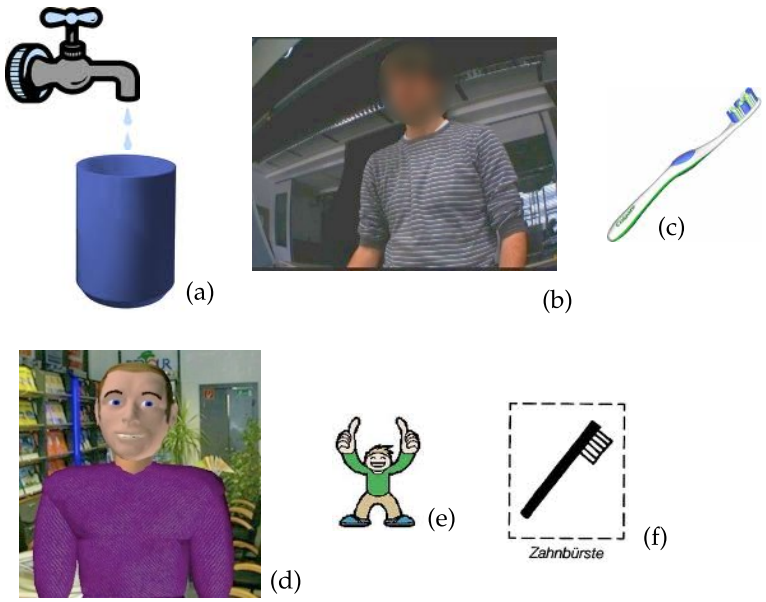


Fig. 5. Selection of prompts which were presented to the caregivers in the interviews. (a) cartoon-like image, (b) frame from a real-life video, (c) real-life image, (d) virtual agent, (e) cartoon-like character, (f) pictogram.

The qualitative analysis of the recorded interviews revealed that an audio command is necessary to attract the user's attention. A visual cue only is likely to fail since the user might miss the visual cue. All caregivers favored short commands in which the textual information is reduced to a minimum. For example, *"Clean mug."* is preferred to *"Please, clean the mug in front of you."* since the shorter command is less cognitively demanding than a longer one. Furthermore, a male voice is preferred to a female voice according to the caregivers. It is negligible whether the voice is known or unknown: according to the caregivers, the effect of an unknown voice will be obsolete after a few trials with the system. The caregivers argued that a verbal command should be accompanied by a visual cue. Two types of prompts were favored: *pictogram* prompts and *real-life videos* showing the desired behavior. Pictograms are most likely to suit most of the user's abilities since users are already familiar with pictogram prompts: such prompts are already part of a user's daily routine in *Haus Bersaba*. However, some users might not be able to understand pictogram prompts, but need a more sophisticated visualization: real-life videos showing the desired behavior are appropriate

for such users according to the caregivers. Two of three caregivers perceived an embodiment of audio commands such as a virtual agent and a cartoon-like character as inappropriate since the characters attract the attention of the users, but do not provide a visual cue of the desired behavior. Additionally, users might feel infantilized by a cartoon-like character. We incorporated the results of the interviews in the development of a two-level prompting hierarchy. Graded prompting hierarchies are a common way to foster a user's independence in task execution by increasing the specificity of prompts during assistance [Demchak 1990]. On the first level, we present pictogram prompts paired with an audio command. If the user doesn't react to a prompt, the TEBRA system will escalate in the prompting hierarchy. On the second level, we present a real-life video of the desired behavior paired with an audio command. Figure

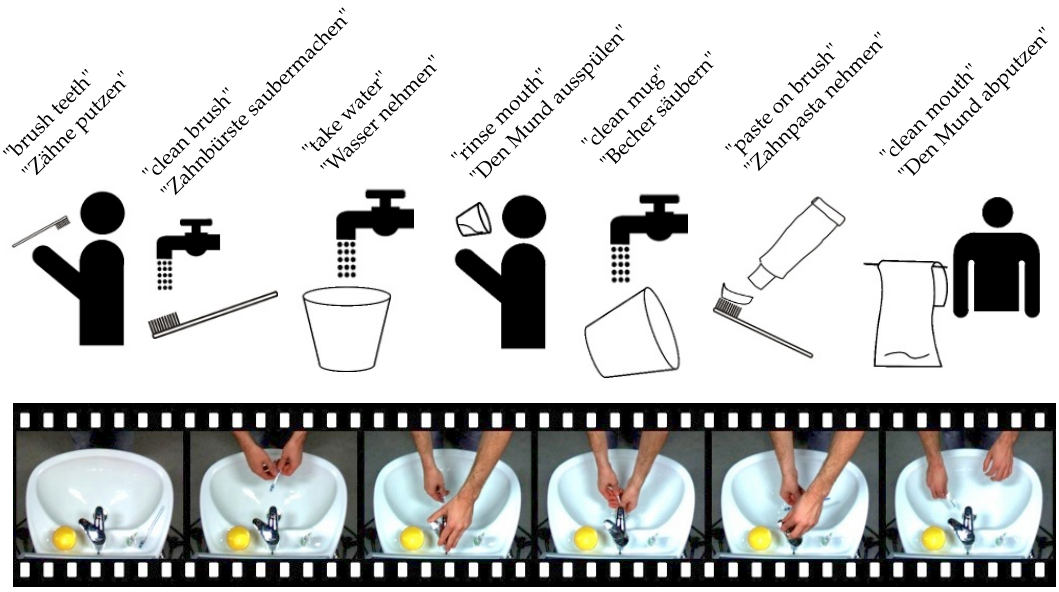


Fig. 6. Pictogram prompts and exact wording of audio commands in English and German. The filmstrip visualization at the bottom exemplarily shows the real-life video of *clean_brush* in single images.

6 shows the exact wording of the audio commands in German and the corresponding translation in English. Furthermore, the pictogram prompts used in the TEBRA system are shown and, exemplarily, the real-life video of *clean_brush* using a film strip visualization. For a single behavior, the same audio command is used in the pictogram prompt and the real-life video prompt: according to the caregivers participating in the interview study, adding more detailed information in the audio command used with the real-life video prompts would most likely distract the users due to the high cognitive demand on the visual and verbal cue.

4. TEBRA SYSTEM

Figure 7 gives an overview of the functional components of the TEBRA system. In this article, we will briefly describe the two main components of the system which are the *Behavior Recognition* and the *Planning and Decision Making* component. For a detailed description of the *Behavior Recognition* component, please refer to the paper [Peters et al. 2012]. The *Planning and Decision Making* component was introduced in [Peters et al. 2013].

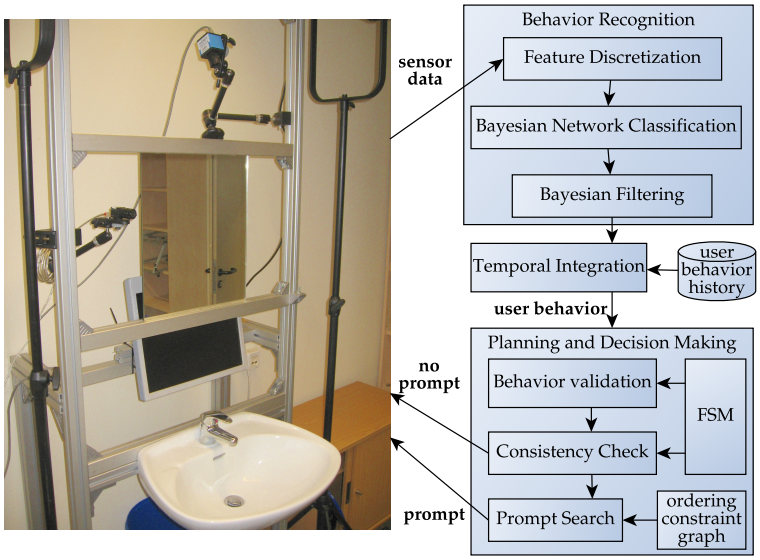


Fig. 7. Functional overview of the TEBRA system.

4.1. Behavior recognition

User behavior recognition is challenging due to the spatial variance in the execution of the task: firstly, a user shows an individual way of performing single behaviors. For example, one user may take the paste with the left hand while spreading the paste on the brush. Another user might use the right hand which results in completely different movement characteristics. Additionally, recognizing behaviors of people with cognitive disabilities is challenging since cognitive disabilities might coincide with motor impairments leading to an even more individualized execution of behaviors [Kluger et al. 1997]. We abstract from recognizing specific movements by tracking objects or the user's hands due to the variance in execution. Instead, we infer a user's behavior based on the environmental configuration encoded in the *behavior* variables *mug_position*, *towel_position*, *tap_condition*, *brush_movement* and *paste_movement* as given in table II which we use as an intermediate representation in our recognition component. The variables *mug_position*, *towel_position* and *paste_movement* are calculated using computer vision techniques on the camera images. We apply a color-based object detector which provides a bounding box hypothesis about the location of an object. The detector is based on a color distribution model of the object which is learned based on sample images of objects. For a detailed description of the color detector, we refer to [Siepmann et al. 2012].

Figure 8 (a) depicts detector results for the mug, towel and paste location in terms of bounding box hypotheses. We compare the center position (x, y) of the best hypothesis of an object to a set of predefined, static regions depicted figure 8 (b). Important regions in the brushing scenario are extracted from the IU analysis results. We identified the *counter*, *hook*, *tap*, *face*, and *else* region denoted with a-e in figure 8 (b) *Hook* denotes the region where the towel is placed when it's not being used. For example in figure 8, the *mug_position*, *towel_position* will be set to *face* and *hook*, respectively. Movement of the paste is discretized into the two values *yes* and *no*. We assume that the paste is placed on the counter unless the user applies the paste. Hence, if the center point of the best hypothesis for the paste is located in the counter region, *paste_movement* will be set to *no*, and otherwise, to *yes*. The condition of the tap will be set according

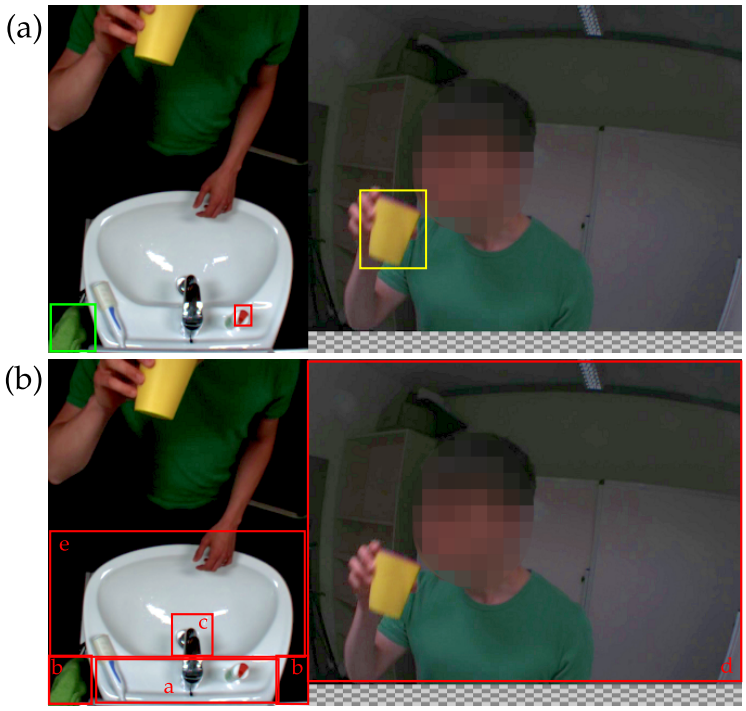


Fig. 8. (a) Bounding box hypotheses for mug and towel. (b) Predefined, static image regions used in the discretization of features. a - counter, b - hook, c - tap, d - face, e - else.

to the flow sensor: if the flow sensor returns 0, *tap_condition* is set to *off*, otherwise to *on*. We apply the gyroscope data and Euler angles provided by the sensor module in the brush to estimate the movement of the brush: the gyroscope measures the angular velocity of the change in orientation. *Brush_movement* will be set to *no*, if the angular velocity is below a threshold over three consecutive time steps. In order to distinguish between *paste_on_brush* and *brush_teeth* (the brush is moving in both behaviors), we use Euler angles which measure the relative orientation of the brush. *Yes_sink* refers to the case when the brush is oriented towards the mirror of the washstand as is usually done in *paste_on_brush*. For *yes_face*, the brush is oriented towards the user which is characteristic for *brush_teeth*. We will set *brush_movement* to *yes_sink* if the orientation of the z component of the brush is $g_z \geq -90$ and $g_z \leq 60$ as illustrated in figure 9. Otherwise, we set *brush_movement* to *yes_face*. We determined the threshold values based on test trials where we evaluated different parameter values. We use a calibration routine prior to a trial which sets the zero orientation according to a fixed initial orientation of the brush to ensure that the zero point of the orientation is persistent over all trials.

IU analysis decomposes the brushing task into user behaviors as given in column *UB* of table I. We subsume the user behaviors *fill_mug* and *clean_mug* to a common user behavior *rinse_mug* in the recognition component because the *behavior* variables involved as well as the according object states are nearly identical for both user behaviors: the mug is given to the tap and the water is turned on. The distinction between filling and cleaning the mug is not observable with the computer-vision techniques used in the TEBRA system. However, we need to distinguish between *fill_mug* and *clean_mug* in the planning and decision making component in order to properly track

Fig. 9. Distinction between *yes_sink* and *yes_face* for variable *brush_movement*. In this example, the orientation of the brush (dashed line) points into the sink region and *brush_movement* is set to *yes_sink*.

the user's progress in the task. In a regular trial of brushing teeth, user behaviors don't follow exactly on each other, but mostly alternate with transition behaviors: for example, a user's hands approach or leave a manipulated object. We consider these transition behaviors by adding a user behavior *nothing* which we treat as any other behavior in our recognition model. We infer a user's current behavior based on the discretized *behavior* variables using a Bayesian network (BN). A BN is a probabilistic graphical model representing a joint probability distribution of random variables. We apply a BN with *Naive Bayes* structure where each *behavior* variable O_i is conditionally independent given the user behavior S:

$$P(o_1, ..., o_5, s) = \prod_{i=1}^{5} P(o_i|s) \cdot P(s) \tag{1}$$

The BN with *Naive Bayes* structure has the ability to deal with small training sets since the probability of each o_i depends only on the user behavior s. This is important in our work, because some user behaviors like *clean_brush* are rare compared to other behaviors. Hence, the amount of available training data is limited. For a detailed description of the behavior recognition, we refer to the paper [Peters et al. 2012].

4.2. Planning and Decision Making

In the behavior recognition component, we can't distinguish between *rinse_mouth_clean* and *rinse_mouth_wet* because the *behavior* variables are nearly identical for both behaviors. Hence, we subsumed the behaviors *rinse_mouth_clean* and *rinse_mouth_wet* to a common behavior *rinse_mouth*. In order to track a user's progress in the overall task properly, we need to distinguish between *rinse_mouth_clean* and *rinse_mouth_wet* since the behaviors have different semantics in the course of the task: *rinse_mouth_wet* describes taking water using the mug before brushing teeth. *rinse_mouth_clean* denotes removing the foam after brushing by rinsing the mouth with water. Furthermore, the behaviors are different in terms of preconditions and effects as given in table III: *rinse_mouth_clean* has the preconditions *mug_content=water*, *mouth_condition=foam* and an additional precondition *teeth_condition=clean*. The preconditions *mouth_condition=foam* and *teeth_condition=clean* can only be provided by the behavior *brush_teeth*. Hence, *brush_teeth* serves as a logical border between the behaviors *rinse_mouth_wet* and *rinse_mouth_clean* during task execution. We use this

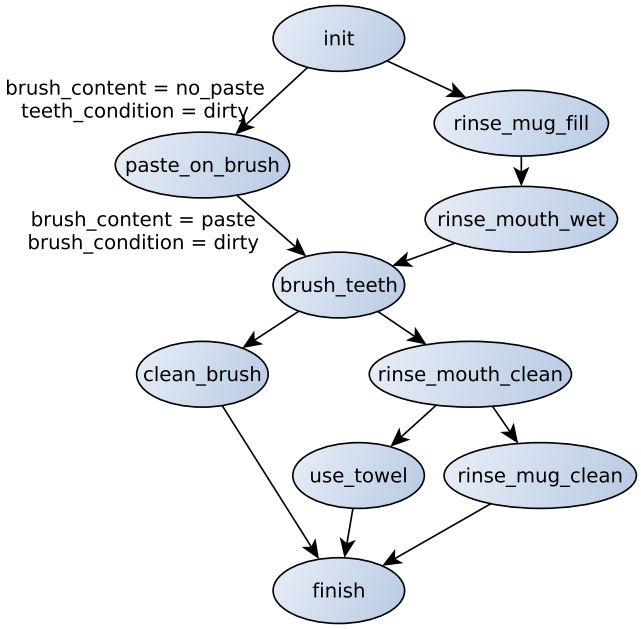


Fig. 10. Ordering constraint graph depicting partial orderings of user behaviors in the brushing task. We depict the preconditions and effects of *paste_on_brush*, exemplarily.

fact in a heuristic in order to distinguish between these behaviors: when *rinse_mouth* is classified by the recognition component, it will be set to *rinse_mouth_wet* if *brush_teeth* has already been recognized during the trial. Otherwise, *rinse_mouth* will be set to *rinse_mouth_clean*. We apply the same heuristic in order to distinguish between *rinse_mug_fill* (when *brush_teeth* has not been recognized) and *rinse_mug_clean* (when *brush_teeth* has already been recognized) which are subsumed to a common behavior *rinse_mug* in the recognition component due to similarities in the involved *behavior* variables.

In case of an inconsistent behavior during task execution, the TEBRA system delivers a prompt to the user indicating the correct behavior. A planning component decides whether a user's progress in the task is consistent as described in the following subsection.

4.2.1. Partial-order planning. We maintain an *ordering constraint graph (OCG)* which models a set of ordering constraints between user behaviors in the overall brushing task. An ordering constraint is a temporal relation $a \prec b$ where a and b are actions and $\prec$ denotes that a precedes b. We calculate the OCG for the tooth brushing task based on a *partial-order planner*. In the TEBRA system, we use the results obtained in the IU analysis to specify the planning domain for the tooth brushing task. The user behaviors and according preconditions and effects as given in table III form the set of actions A. The initial state I and the goal state G are extracted from the IU analysis in table I. We manually constructed the OCG as depicted in figure 10 from the results of the partial-order planner. An arrow in the OCG describes that the source behavior provides necessary preconditions for the target behavior. For example, *rinse_mug_fill* provides the effect *mug_content=water* which is a precondition of *rinse_mouth_wet*. The OCG depicts no strict execution plan of the task which the user has to follow, but

models the ordering between behaviors in the overall task: for example, the behavior sequence *rinse_mug_fill*, *paste_on_brush*, *rinse_mug_fill* is consistent with respect to the partial ordering given in figure 10. Modeling the partial ordering is desirable in the TEBRA system since it allows a user to perform the brushing task in an individual way as long as the overall constraints represented in the OCG are met during task execution. Furthermore, the OCG representation is much more compact with regard to memory consumption in comparison to modeling any allowed transition from the initial state to the goal state explicitly.

A successful execution of the tooth brushing task is a transition from an initial state to a final state in the progress state space. Transitions between states are triggered based on the occurrence of user behaviors: we are not able to robustly recognize whether the effects of user behaviors have occurred due to the limited capabilities of the sensor technology. For example, it is nearly impossible to detect whether a user has spread paste on the brush based on computer vision techniques. Furthermore, an additional sensor for this purpose is not desirable due to hygienic reasons. Hence, we infer the occurrence of behavior effects based on the duration of behaviors. When a behavior is recognized over a certain period of time, we infer that the user has successfully performed the behavior and update the progress state space with the effects of the behavior. An appropriate update is challenging with regard to the temporal variance in the execution of behaviors due to (1) different durations of behaviors and (2) different velocities of users in task performance. In the following subsection, we will describe a dynamic timing model which is able to handle the temporal variance.

4.2.2. Dynamic timing model. We explicitly model the timing characteristics of user behaviors in a *dynamic timing model* to track a user's progress in the task properly with regard to the following principle: we aim to prevent a user from performing an erroneous behavior by checking the consistency of the behavior as early as possible. If the consistency check is too late, the behavior effects might have erroneously occurred already. This might lead to an inconsistent state space and erroneous prompts during the remainder of the task.

We subdivide user behaviors into three phases: validation, pre-effect and post-effect. Transitions between two phases denote important events in the planning and decision making component. At the transition from the validation phase to the pre-effect phase, we check the consistency of the current user behavior with regard to the progress state space after a validation time t_v. The duration of the validation phase ensures that a user's current behavior is persistent over a period of time. Hence, we avoid delivering erroneous prompts due to temporary errors in the recognition component. At the transition from the pre-effect to the post-effect phase, we update the progress state space with the effects of the current behavior after an effect time t_e. For any user behavior, a *timeout* t_t^s may occur in the *post_effect* phase. A timeout denotes that the user might not be able to terminate the behavior, e.g. due a user's obsessiveness in task execution. We model the phases of behaviors using a Finite State Machine (FSM). For a detailed description of the FSM and the exact calculation of the timing parameters, we refer to [Peters et al. 2013].

In order to cope with the variance in the duration of individual behaviors, we maintain a timing model $\mathbf{t^s} = (t_v^s, t_e^s, t_t^s)$ for each user behavior s. For example, the duration of *use_towel* is usually much shorter compared to *brush_teeth*. Hence, the effect time t_e^s and timeout t_t^s of the behaviors are completely different. The validation time t_v^s can be set higher for longer behaviors to avoid a misdetection of the behavior due to perception errors.

In addition to different durations of user behaviors, users show different velocities in the execution of behaviors due to individual abilities. In the TEBRA system, we al-

low for different user velocities by maintaining timing models for three different user velocities corresponding to *fast*, *medium* and *slow* execution velocity. The three velocity categories were chosen manually by the authors based on the in-situ observations described in section 3.1.

Table IV gives an overview of the timing parameters in seconds. We manually adjusted the timing parameters in two ways: firstly, we set a minimum time for behavior *brush_teeth* proposed by the caregivers in order to ensure that the teeth are sufficiently cleaned. Hence, we set the effect time $t_e^s = 60s$ for behavior $s =$ *brush_teeth* in each velocity model. Secondly, we check the consistency of a user behavior after a maximum behavior duration of $5s$ in order to prevent a user from performing an inconsistent behavior over a long period of time. Hence, we set the validation time, after which a consistency check is triggered, to $t_v^s = max(t_v^s, 5)$ for each behavior s in each velocity model. The adjustments of the validation time affected behavior *paste_on_brush* in velocity *slow* and *brush_teeth* in velocities *medium* and *slow*. We apply the learned timing parameters in a dynamic timing model which chooses the timing parameters of the FSM according to the user's current velocity in a trial. When the user terminates a behavior s, we determine the duration t_s. We categorize the duration into one of the velocity classes *fast*, *medium* and *slow* using the probability density functions of the Gaussian distributions of behavior s (see [Peters et al. 2013] for details). The velocity class of behavior s is the class that has most likely produced the behavior with the current duration. During a trial, we count the number of occurrences of behaviors of each velocity class. We set a user's current velocity by applying a winner-takes-all method on the velocity counts which chooses the velocity occurring most frequently during the trial so far. In the beginning of a trial, we don't use prior knowledge about a user's velocity in former trials. Hence, we allow for differences in a user's velocity between trials which might arise due to daily mood or effects of temporary medication.

Table IV. Parameters of the dynamic timing model in seconds for user behaviors in the different velocities. t_v, t_t and t_e - validation, timeout and effect time.

User behavior	**fast**			**medium**			**slow**		
	t_v	t_e	t_t	t_v	t_e	t_t	t_v	t_e	t_t
paste_on_brush	1.4	3.7	17.5	3.4	10.3	35.8	5.0	24.0	60.5
rinse_mug_fill	0.5	1.6	6.4	1.1	3.3	11.0	2.3	7.3	21.9
rinse_mug_clean	0.6	1.9	6.8	1.2	3.5	12.0	2.4	7.3	22.9
rinse_mouth_wet	0.4	1.4	4.4	0.7	2.0	6.2	1.0	3.1	9.2
rinse_mouth_clean	0.5	1.6	5.6	0.9	2.6	8.7	2.2	6.1	25.0
brush_teeth	3.1	60.0	55.7	5.0	60.0	194.7	5.0	60.0	426.5
clean_brush	0.5	1.4	6.6	1.7	5.0	18.6	4.5	11.8	56.0
use_towel	0.8	2.3	10.3	1.7	5.1	17.7	3.1	9.7	30.0

4.2.3. Prompt selection. We select an appropriate prompt using a search procedure in the OCG. We determine the open preconditions of the inconsistent user behavior s. We search for a user behavior s' which is a predecessor of s in the OCG and provides at least one open precondition. If s' exists, we check the consistency with regard to the progress state space. When s' is consistent, s' is an appropriate prompt. If s' is also inconsistent due to open preconditions, we recursively search for a behavior resolving the open preconditions of s'. Hence, we are able to resolve chains of open preconditions over several user behaviors by iterating backwards through the OCG. If no predecessor of s is found providing the open precondition, we search for a consistent behavior by iterating backwards through the OCG starting at the *finish* node. By starting at the finish node, we aim to find a consistent behavior which is most closely to the desired goal state. Furthermore, we avoid prompting for a behavior which the user has already

performed or which doesn't yield progress in the overall task. In case of a timeout, a user's current behavior is consistent without open preconditions since the behavior has already passed the consistency check during performance. Hence, the prompt selection mechanism directly searches for a consistent follow-up behavior starting at the *finish* node.

5. USER STUDY

The study with people with cognitive disabilities described in this section is the first study where we deploy a prototype of the TEBRA system to target group users. We cooperate with the residential home *Haus Bersaba* belonging to the *v. Bodelschwinghsche Stiftungen Bethel*, a clerical foundation in Bielefeld. 35 people with cognitive disabilities live permanently in *Haus Bersaba* and receive professional nursing care in their everyday life.

The recruiting of participants - called users in the following - was based on inclusion and exclusion criteria which we assessed in conjunction with the caregivers of *Haus Bersaba*. We included users who (1) are motivated to participate in the study, (2) are reliant on a caregiver for successful execution of the tooth brushing task, (3) show appropriate perception and responsiveness to react to verbal and visual assistance, (4) are aged between 18 and 75 and have an IQ greater than 35. Exclusion criteria were severe physical disabilities which prevent the user from fulfilling the task. For example, a user needs to have the motor skills to hold and use the toothbrush. Furthermore, a decreased ability in visual perception which prevents a user from perceiving prompts on the screen, as well as serious medical conditions such as heart deficiency and cancer are exclusion criteria.

The data recorded during the study is sensitive with regard to privacy concerns: we record data with different sensors including cameras which show users in tooth brushing which is a private activity in a user's bathroom. All participants in the study (caregivers and users/legal guardians) signed a declaration of consent and a sheet of information where we described the study procedure as well as the privacy policy. The privacy policy includes that we (1) treat the acquired data strictly confidentially, (2) restrict the data access to the investigators of the study, and (3) anonymize the data prior to evaluation. Furthermore, a user is able to terminate the participation in the study at any time without giving any reasons. In order to ensure the appropriateness of the study with regard to privacy issues as well as ethical and nursing aspects, we applied for ethical approval at the ethics committee of *Westfälische Wilhelms-Universität Münster*. The ethics committee approved the application without limitation.

5.1. Study design

The group of participants in our study consists of seven users. Table V shows demographic information about the participants. All participants have an IQ greater than

Table V. Demographic information about the study participants.

user	gender	age	disabilities
1	m	41	intellectual disabilities, autistic spectrum disorder, epilepsy
2	f	56	intellectual disabilities
3	f	53	behavioral disorder, intellectual disabilities
4	f	45	autistic spectrum disorder, intellectual disabilities
5	m	56	intellectual disabilities, epilepsy
6	m	48	behavioral disorder, intellectual disabilities
7	f	55	intellectual disabilities

35. The exact IQs of individual participants are not known to the authors. All partic-

ipants have a working knowledge of all objects used in the task. The target group is heterogeneous since the users have different types of moderate cognitive disabilities. Due to the heterogeneous user group and the small sample size of seven users, general hypotheses in terms of diagnostic assessment and therapeutic treatment of users with specific disabilities are not feasible. Instead, we evaluate the influence of the TEBRA system on a user's individual behavior in brushing teeth.

We follow a *single-subject design* approach widely used in behavioral science [Richards et al. 1998; Robson 2002]. We evaluated the user's behavior in an AB study design where A and B correspond to the baseline and intervention phase, respectively. The treatment variable here is the entity that provides a user's assistance which is either the caregiver or the TEBRA system. In the caregiver (CG) scenario (baseline phase), users brush their teeth at the washstand. The TEBRA system is working in a way that sensor data is recorded and the user's overall progress in the task is tracked, but the delivery of prompts is suppressed. Instead, a caregiver standing besides the washstand, assists the user in the brushing task. The CG scenario is the regular way of task assistance in *Haus Bersaba* since all users in our study are reliant on the assistance of a caregiver in brushing teeth during their daily routine. In the system (SYS) scenario (intervention phase), users are assisted by the TEBRA system which provides audio-visual prompts via the display installed at the washstand. A caregiver, who is hidden behind a room divider, is present in each SYS trial in order to intervene and take over the assistance in case of fatal system errors.

The seven users conducted trials on nine different days. Each user performed only a single trial in the recording session of a day. We ensured that the trials smoothly integrate into a user's daily routine in order to evaluate the user's behaviors in regular situations as far as possible. Hence, we aimed to align the study times with the regular tooth brushing times of the users by conducting the trials in the evenings. We recorded a total of 55 trials: 20 in the CG scenario and 35 in the SYS scenario. One user skipped five trials (1 CG, 4 SYS) due to motivational reasons and participated only in two CG and SYS trials, each. Two trials of user 2 and a single trial of user 3 were terminated due to a system crash and the caregiver assisted the users in the remainder of the task. In CG, the same caregiver assisted in each of the 20 trials.

The main aim of the study is to analyze the user's behavior in the interaction with the system with regard to the main hypothesis: Is the TEBRA system able to support the independence of users in the execution of brushing teeth? Our study results provide evidence that this support is being provided by the TEBRA system.

We present and discuss the results of the study in the following section.

5.2. Results and discussion

In order to assess quantitative results, we segmented the trial data into the behaviors given in table IV. We followed a systematic coding scheme using a conjunction of events which describe the beginning and the end of a behavior, respectively. Table VI gives an overview of the segmentation methodology: for example, the beginning and end of behavior *paste_on_brush* is determined using the movement of the paste. When the paste dispenser leaves the counter region, *paste_on_brush* starts. The behavior ends when the paste dispenser enters the counter region after the paste was taken. The segmentation was manually done by the first author of the paper. Given the precise nature of the coding scheme, it is very unlikely that a different coder would result in any substantial differences.

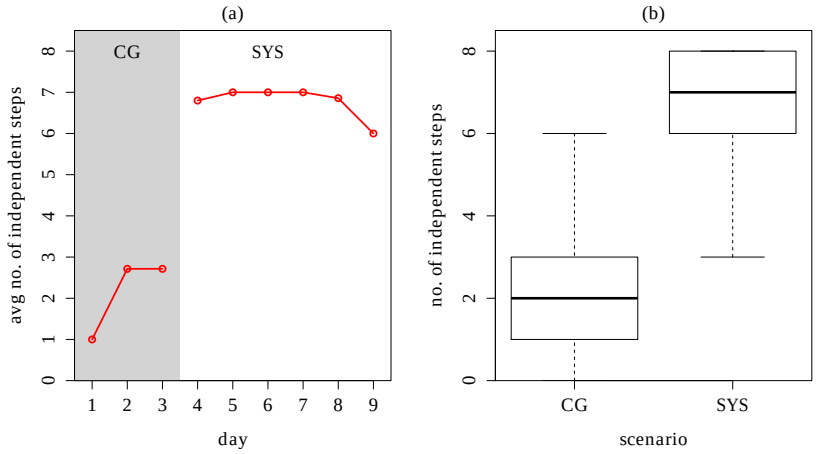
The TEBRA system aims to increase the independence of users and improve their self-confidence by providing appropriate assistance in task execution. An important measure of the influence of the TEBRA system is the number of independent steps - the number of steps which a user is able to perform without the help of a caregiver. For

Table VI. Segmentation methodology used to annotate the user behaviors.

User behavior	Start events	End events
paste_on_brush	paste leaves counter region	paste enters counter region
rinse_mug_fill	mug enters tap region	mug leaves tap region
	water running	*brush_teeth* not done, yet
	brush_teeth not done, yet	
rinse_mug_clean	mug enters tap region	mug leaves tap region
	water running	*brush_teeth* already done
	brush_teeth already done	
rinse_mouth_wet	mug enters frontal image	mug leaves frontal image
	brush_teeth not done, yet	*brush_teeth* not done, yet
rinse_mouth_clean	mug enters frontal image	mug leaves frontal image
	brush_teeth already done	*brush_teeth* already done
brush_teeth	brush moving	brush moving
	brush oriented towards the user	brush oriented towards the mirror
clean_brush	brush enters tap region	brush leaves tap region
	water running	
use_towel	towel enters frontal image	towel leaves frontal image

example, a user adapting his/her behavior due to a system prompt is an independent step of the user since no caregiver is involved in prompting. In the SYS scenario, the number of independent steps is significantly increased compared to the CG scenario. Figure 11 (a) shows the average number of independent steps on the nine days of the study where the CG scenario comprised three and the SYS scenario six study days. The average number of independent steps in the CG scenario is stable on days 2 and 3

Fig. 11. (a) Average number of independent steps per trial day. (b) Boxplot of number of independent steps in the CG and SYS scenario. The different steps of the brushing task are listed in table III.



with around 2.7 steps. In total, the caregiver gave 105 prompts during the trials which makes an average of 5.8 prompts per trial. On day 1 of the CG scenario, the average number is very low with 1.0 independent steps only. The users brushed their teeth at the unfamiliar washstand for the first time. According to the caregiver, users were highly excited due to the start of the study and the recording of their performance. Hence, the users were unconcentrated which resulted in a poor performance in terms of the low number of independent steps. The average number in the SYS scenario is

stable over five days with around 7. During the SYS trials, the caregivers had to step in 17 times. The caregiver stands behind a room-divider and follows the performance of the user. We briefed the caregivers to step in at any time if they feel that the user's performance is bad or if the user is confused by the system prompts. In none of the SYS trials, the caregiver stepped in actively. However, the caregiver reacted to the users in situations where a user directly approached the caregiver throughout the trials by asking for help. These situations concentrate mainly on the trials of user 2 who approached the caregiver in 14 of the 17 CG prompts. In 29 trials, the caregiver did not provide any assistance at all. We briefed the caregivers to finalize the brushing task in an case of an insufficient performance. Since the caregiver did not finalize in any of the SYS trials, the users successfully brushed their teeth in all SYS trials. The average result on the last day of the SYS trials is decreased due to a single user's performance: user 6, who completely skipped four SYS trials due to motivational issues, quit the trial after three steps and left the room due to unknown reasons. Up to the time where user 6 left the room, the performance of the system was not overly erroneous. We conclude that the user left due to personal reasons and not due to inappropriate assistance by the TEBRA system. The decreased number of independent steps in this trial decreased the average rate shown in figure 11 (a). In the following, we will drop the results of user 6 due to the limited amount of data available (only two CG trials and a single SYS trial). The visual inspection of the average number of independent steps reveals a significant difference between the CG and the SYS scenario. The statistical significance of the difference is tested using a non-parametric test. We apply a non-parametric, Mann-Whitney U-Test since the average number of independent steps is skewed and, hence, not normally distributed according to figure 11 (b). Based on the empirical data, the test provides a significant result with $U = 16$ and $p = 3.5 \cdot 10^{-9}$. We reject the null hypothesis since the value of $p < 0.05$. We infer that the application of the TEBRA system has an effect in terms of an increased average number of independent steps of users.

The average results over all users hide variations between individual users. Figure 12 shows the number of independent steps for individual users over trials. A red cross denotes a trial in which the system crashed due to technical problems with the Bluetooth connection of the brush which occurred in three SYS trials. Users 3 and 4 show excellent results using the TEBRA system: all trials of user 4 were perfect trials in a way that all eight sub steps of the task were performed independently of a caregiver. User 3 has similar results with an average number of 7.8 independent steps per trial. In comparison to users 3 and 4, user 2, for example, has a lower number of independent steps with 5.5 per trial. In the last trial of user 2, the number of independent steps drops from about five or six independent steps in the previous SYS trials to three: in this trial, user 2 wore a yellow shirt which was very similar in color appearance compared to the yellow mug used in the trials. Parts of the yellow shirt were erroneously recognized as the mug on the frontal image. Hence, the discretization of the mug detector hypothesis into the position of the mug was error-prone throughout the whole trial. This resulted in errors in the classification of user behaviors and, hence, to an increased number of false prompts during the course of the trial. The false prompts confused user 2 in task execution which led to the decreased number of three independent steps in this trial. All users show an increase in the number of independent steps from the CG to the SYS scenario. The amount of increase varies between individual users as shown in table VII. User 7 shows the best performance in the CG scenario amongst all users with 4.7 independent steps on average. However, the increase of independent steps from the CG to the SYS scenario is low with 1.6. The benefit of the TEBRA system is quantitatively lower for user 7 compared to other users. However, the quantitative increase of 1.6 might be clinically meaningful for the user and the

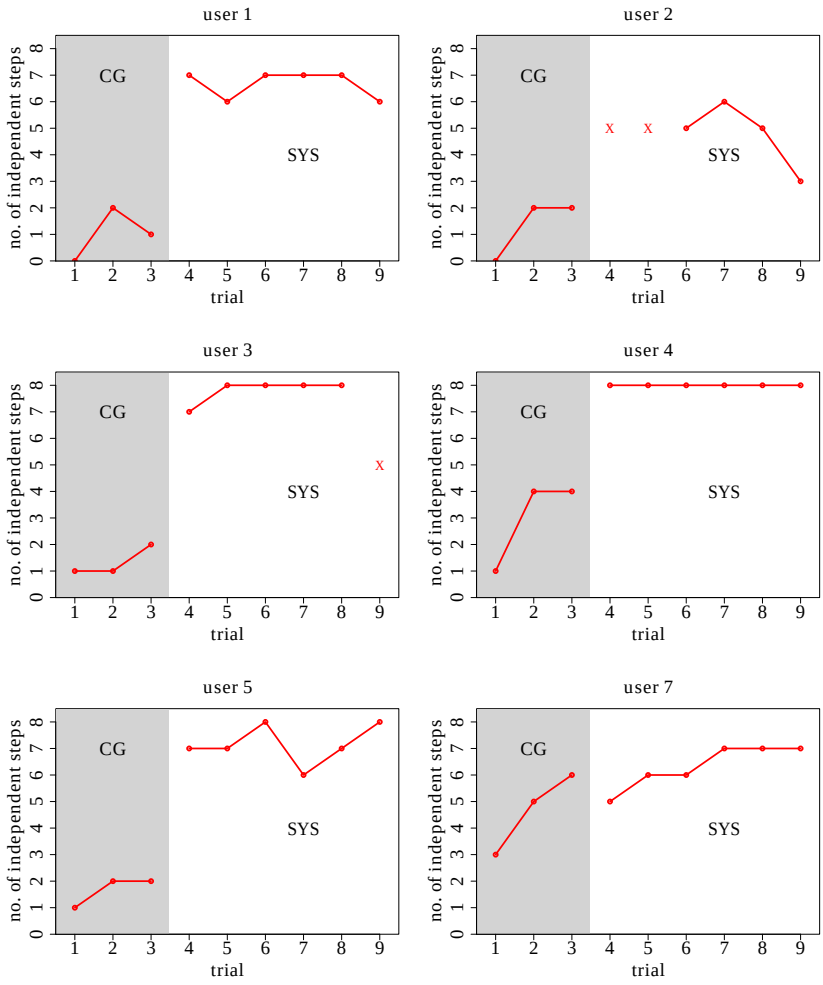


Fig. 12. Number of independent steps per trial for individual users.

Table VII. Average number of independent steps in the CG and SYS scenario for individual users and increase in the number of independent steps from CG to SYS.

		user					
		1	2	3	4	5	7
avg no. of ind. steps	CG	1	1.3	1.3	3	1.7	4.7
	SYS	6.7	4.8	7.8	8	7.2	6.3
	CG to SYS	+5.7	+3.5	+6.5	+5	+5.5	+1.6

caregiver. User 3 showed an increase of 6.5 independent steps from the CG to the SYS scenario which is an outstanding increase. Hence, the benefit of using the TEBRA system is great for user 3 who showed a low average of 1.3 independent steps in the CG scenario. In the following subsections, we further analyze the overall performance by

Table VIII. Classification rates of user behaviors in the SYS scenario in %. RMgC - rinse_mug_clean, RMgF - rinse_mug_fill, UT - use_towel, PB - paste_on_brush, RMC - rinse_mouth_clean, RMW - rinse_mouth_wet, BT - brush_teeth, CB - clean_brush, N - nothing.

	RMW	**RMC**	**RMgF**	**RMgC**	**BT**	**PB**	**CB**	**UT**	**N**
RMW	**82.5**	0.0	4.8	0.0	0.0	1.2	0.6	0.0	10.9
RMC	31.6	**41.0**	1.1	5.3	0.0	0.9	0.1	2.6	17.4
RMgF	1.3	1.9	**54.9**	24.2	0.0	0.0	0.5	1.9	15.1
RMgC	0.0	0.0	1.2	**80.4**	0.0	2.3	8.1	3.0	5.0
BT	7.0	0.8	1.3	0.4	**50.1**	25.1	2.7	0.3	12.3
PB	0.3	0.0	0.0	0.0	1.5	**97.8**	0.0	0.0	0.4
CB	0.0	0.3	0.0	4.3	0.0	3.2	**79.9**	10.6	1.7
UT	0.0	2.2	0.0	3.3	0.0	2.2	2.5	**75.5**	14.3
N	4.6	3.0	1.6	2.4	4.4	13.2	4.1	4.7	**62.0**

evaluating the recognition component and the TEBRA system's ability to deal with spatial and temporal variance in task execution.

5.2.1. Technical evaluation. A key challenge for providing appropriate prompting is the recognition of user behaviors. The major challenge in behavior recognition is the spatial variance in task execution: spatial variance describes the different movement characteristics of individual users during behavior execution. For example in the execution of *clean_brush*, one user was holding the tap while cleaning the brush. Another user cleaned the whole brush under the tap. Furthermore, the user's hands are partly or fully occluded. A recognition using a hand or an object tracker would not be feasible due to occlusions. We abstract from the recognition of movement trajectories of objects or the user's hands, but instead infer user behaviors based on states of objects involved in the behaviors. Table VIII shows the results of the user behavior recognition component for the trials in the SYS scenario. The average recognition result over all user behaviors is 69.3%. The classification rates of single behaviors vary between 97.8% for *paste_on_brush* and 41% for *rinse_mouth_clean*. The results of the behaviors *rinse_mouth_wet*, *rinse_mug_clean*, *clean_brush* and *use_towel* range from 75.5% to 82.5%. These results are good with regard to the spatial variance in task execution. However, the rates of *rinse_mouth_clean*, *rinse_mug_fill* and *brush_teeth* are poor with 41%, 54.9% and 50.1%, respectively. *Brush_teeth* is mixed up with *paste_on_brush* in 25.1% of the cases. Obviously, the classification based on the orientation of the brush is error-prone. The classification rates of *brush_teeth* vary extremely between different users. For example, user 3 has an average classification rate of 97% for *brush_teeth* over all SYS trials. The average classification rate of user 1 is only 9% for *brush_teeth*. User 1 leans over heavily while brushing teeth. The brush is oriented in a way that the discretization of the *brush_movement* is set to *yes_sink* instead of *yes_face* which leads to a misclassification of *brush_teeth* as *paste_on_brush*. Hence, the poor classification rates of specific users decrease the overall recognition rate of *brush_teeth* to 50.1%. The recognition rate of *brush_teeth* influences the recognition rates of *rinse_mouth_clean* and *rinse_mug_fill* which are 41% and 54.9%, respectively. *Rinse_mouth_clean* was misclassified as *rinse_mouth_wet* with 31.6%. The misclassification concentrates on trials in which the recognition rate of *brush_teeth* is poor: *rinse_mouth_clean*, which is performed after *brush_teeth*, is classified as *rinse_mouth_wet* because *brush_teeth* was not recognized properly. *Rinse_mug_fill* is misclassified as *rinse_mug_clean* with 24.2%. The misclassification mainly concentrates on trials in which *brush_teeth* was properly recognized. Users tended to wet their mouth prior to *brush_teeth* until no water was left in the mug due to obsessive behavior. When they aimed to perform *rinse_mouth_clean* after a successful execution of *brush_teeth*, they started to fill the mug with water again.

Hence, a regular *rinse_mug_fill* behavior was misclassified as *rinse_mug_clean* since the heuristic doesn't model these situations. The classification results show that the recognition component used in the TEBRA system is able to deal with variances in spatial task execution for most behaviors in the brushing task. However, we aim to improve the overall recognition rates by improving two aspects: firstly, the recognition rates of behaviors are highly dependent on the rate of *brush_teeth*. Hence, the improvement of recognizing *brush_teeth* is very important for a successful user behavior recognition in the overall task. Secondly, we need to improve the heuristic which discriminates between *rinse_mouth_wet* (*rinse_mug_fill*) and *rinse_mouth_clean* (*rinse_mug_clean*) in order to avoid misclassifications due to modeling errors.

In addition to the spatial variance, temporal variance is expressed in both inter-behavior and intra-behavior timing differences: inter-behavior differences are variations in the duration of behaviors amongst each other. Table IX gives an overview of average durations of behaviors for all SYS trials. The average duration of individ-

Table IX. Minimum, maximum and average duration of user behaviors.

User behavior	Durations in sec.		
	avg	min	max
paste_on_brush	9.8	2.8	28.4
rinse_mug_fill	2.5	0.5	9.5
rinse_mug_clean	3.2	0.8	7.9
rinse_mouth_wet	2.5	0.9	9.5
rinse_mouth_clean	2.4	0.6	8.6
brush_teeth	67.9	19.0	143.0
clean_brush	5.2	0.7	16.1
use_towel	12.0	1.8	73.3

ual behaviors in the brushing task ranges from $2.4s$ for *rinse_mouth_clean* to $67.9s$ for *brush_teeth*. As shown with the classification rates of table VIII, the recognition component is able to deal with behaviors varying significantly in duration: for example, the average durations of *paste_on_brush* and *rinse_mouth_wet* are $9.8s$ and $2.5s$, respectively. The classification rate for *rinse_mouth_wet* is very good with 82.5% and excellent for *paste_on_brush* with 97.8%.

The durations vary not only between different behaviors, but also in different executions of a single behavior (called *intra-behavior difference* in the following). Intra-behavior difference arises from different velocities in task execution due to a user's individual abilities. For example, the durations of single executions of *paste_on_brush* range from $2.8s$ to $28.4s$. We apply a dynamic timing model to deal with intra-behavior variations and different velocities of users. We will describe the benefit of the dynamic timing model in two situations. Figure 13 visualizes the state of the FSM (black line), the estimate of the user's behavior according to the recognition component (blue line), the estimate of the user's velocity (thick red line), and the ground truth annotation of behaviors (thin red line). The visualization covers an interval of about six seconds in a trial of user 5. User 5 finishes *paste_on_brush* at about $40.3s$. Due to the duration of *paste_on_brush* and the velocities of the preceding behaviors, the velocity model is updated from *medium* to *fast* at $40.5s$. At $41.8s$, the user starts *rinse_mug_fill* which is performed for $2.2s$. Due to velocity model *fast*, the effects of the behavior occur after $1.6s$ which is depicted by the vertical blue line at $43.4s$. With the model for *medium* velocity, *rinse_mug_fill* would not have been recognized correctly since the effect time of $3.3s$ would not have been reached. Hence, the effects of *rinse_mug_fill* would not have

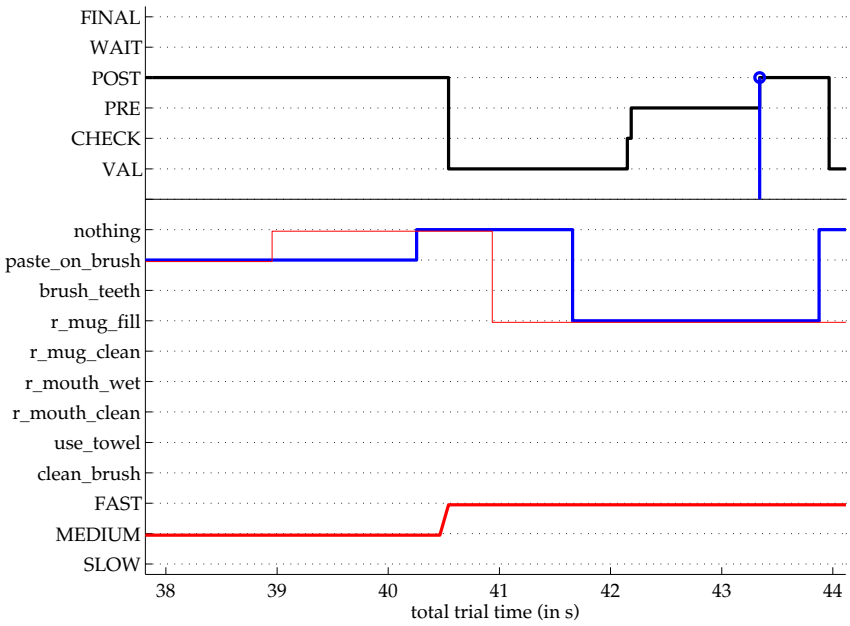


Fig. 13. Section of a trial of user 5 showing the state of the Finite State Machine (black line), the estimate of the user's behavior according to the recognition component (blue line), the estimate of the user's velocity (thick red line), and the ground truth annotation of behaviors (thin red line). The vertical blue line denotes the update of the state space by applying the effects of the current user behavior which is *rinse_mug_fill* here.

been applied to the progress state space leading to erroneous prompts in the remainder of task execution. Figure 14 shows a second situation from a trial of user 2. The user performs *rinse_mug_fill* which is successfully recognized by the TEBRA system. The progress state space is updated with the effects of *rinse_mug_fill* after about $25.6s$ which is depicted by the vertical blue line. The user forgets to perform *rinse_mouth_wet* and *paste_on_brush*, and erroneously starts *brush_teeth* at $32s$. Due to the inconsistency of *brush_teeth*, a pictogram prompt for *rinse_mug_fill* is delivered at about $35s$ which is shown by the vertical black line. The dynamic timing model with velocity *fast* delivers a prompt which is appropriate in time in a way that the user is assisted in the erroneous performance of the task as soon as possible. With a *medium* or *slow* velocity, the prompt would have been delayed and the user would have performed the erroneous behavior for a longer period of time.

A disadvantage of the dynamic timing model is the inclusion of durations of erroneously classified behaviors in determining a user's velocity in a trial. For example, *brush_teeth* is misclassified as *paste_on_brush* for a duration of $3s$. The duration of $3s$ is classified into velocity *fast*. Hence, the dynamic timing model erroneously increases the frequency counter of velocity model *fast* which leads to a skewed distribution of counts over the velocity classes. This might result in a wrong application of timing parameters and the delivery of false prompts in the remainder of the trial. However, as shown in the previous examples, the TEBRA system can deal with intra-behavior variances in temporal execution of behaviors by adapting to the user's velocity during task execution. In the following subsection, we will analyze the prompting behavior of the system and the user's reaction behavior in detail.

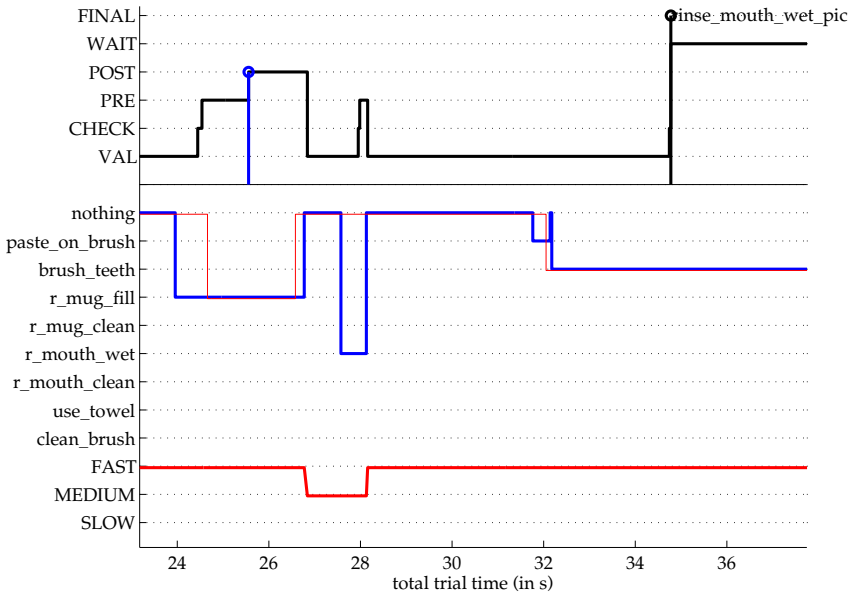


Fig. 14. Section of a trial of user 2. For a description of the lines, see figure 13. The vertical black line denotes that a prompt was triggered by the TEBRA system at that time. Here, a pictogram prompt for behavior *rinse_mouth_wet* was delivered.

5.2.2. Prompting behavior and users' reactions. An important measure for a user's responsiveness to system prompts is the reaction behavior. We classify the reactions of users into three categories: *correct*, *false* and *no* reaction. A user's reaction to a prompt is correct when the user adapts his/her behavior according to the prompt by performing the behavior he/she was prompted for. If the user reacts to the prompt, but does not perform the desired behavior, the reaction will be classified as a false reaction according to the prompt. If the user does not show a reaction at all, we will refer to it as no reaction. In order to further evaluate the appropriateness of prompts, we take into account the number of semantically correct prompts as a measure of appropriateness. Semantically correct means that the type of prompt is appropriate with regard to the user's progress in the task so far. For example, a user has successfully filled the mug with water and gets stuck in task execution. An appropriate prompt in this situation would be either *rinse_mouth_wet* or *paste_on_brush*. We determine the semantic correctness by using a ground truth annotation of the behaviors in the task which was done by the first author of the article. The left plot in figure 15 shows the ratio of semantically correct prompts in the SYS scenario for individual users. The ratio of user 4 is excellent since 93.8% of the prompts are semantically correct. For users 2 and 3, the ratios of semantically correct prompts are good with 82.7% and 81.7%, respectively. However, the percentage for users 5 and 7 are decreased with 57.2% for user 5 and 44.9% for user 7. The low ratios of semantically correct prompts stem from erroneous follow-up prompts due to perception errors in the recognition component: for example, a user performs *rinse_mug_fill*, but the TEBRA system misses the behavior. The user performs *rinse_mouth_wet* subsequently which is a correct behavior according to the course of the trial. However, the system prompts the user to perform *rinse_mug_fill* which is semantically incorrect at that time. If the user does not react to the prompt, the system is likely to issue follow-up prompts for *rinse_mug_fill* which are semantically incorrect, too. The TEBRA system is able to limit the number of erroneous follow-up prompts

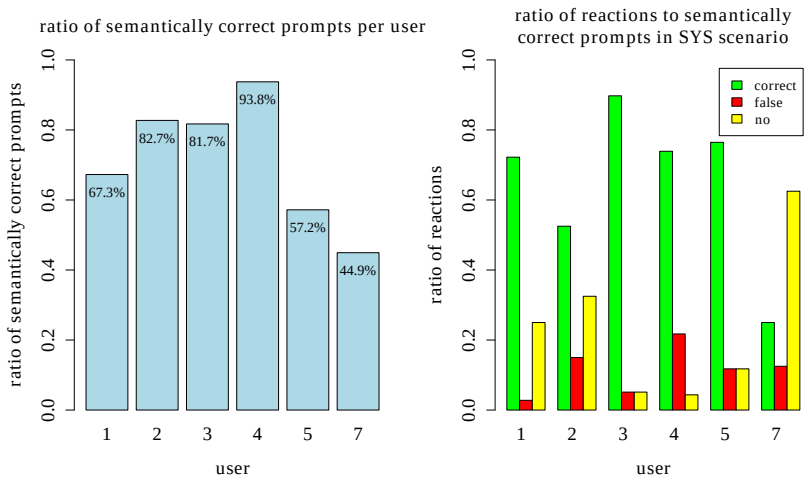


Fig. 15. Left plot: ratio of semantically correct prompts in the SYS scenario for individual users; Right plot: ratio of reactions to semantically correct prompts in the SYS scenario for individual users.

using the following heuristic: after three consecutive prompts of the same behavior (one pictogram and two video prompts according to the escalation hierarchy), the system infers that it has made a perception error and applies the effects of the prompted behavior to the state space. Due to the heuristic, the system is able to recover from perception errors in which a user's behavior was missed during the execution of a trial.

In order to assess a user's responsiveness to prompts, we focus on reactions to semantically correct prompts because the appropriateness of semantically correct prompts is ensured. The right plot in figure 15 shows a user's reactions to semantically correct prompts. Users 3 and 4 show 82% and 75% correct reactions to semantically correct prompts. Users 2 and 7 show only 45% and 20% correct reactions. Instead, the ratio of no reactions to semantically correct prompts is 60% for user 7. Two explanations are possible for the reaction behaviors of users 2 and 7: firstly, they might not be willing to react to the prompts given by the TEBRA system although the prompts are semantically correct. Secondly, they might not be able to understand and react correctly to the majority of system prompts since the presentation of prompts is inappropriate. In the TEBRA system, we use pictogram and real-life videos to prompt the users. We analyze whether pictogram or video prompts are inappropriate for an individual user: figure 16 shows the ratio of correct reactions to semantically correct prompts for pictogram and video prompts. During the analysis of trials with people with cognitive disabilities, we observed that the TEBRA system provides prompts which are consistent with regard to a user's overall progress, but which are not necessary for the user because they were triggered due to perception errors for behaviors with a long duration: for example, the TEBRA system misclassifies *brush_teeth* as *paste_on_brush* prior to the effect time of $60s$ for *brush_teeth*. According to the progress state space, *paste_on_brush* is inconsistent. Hence, a *brush_teeth* prompt is triggered which is consistent with regard to the user's overall progress in the task. Although the prompt occurred due to a perception error and the prompt might not have been necessary for the user since he/she is already performing *brush_teeth*, it is semantically correct with regard to the progress state space: the effect time of the behavior has not been reached and the progress state space has not been updated with the effects of the behavior, yet. We refer to such prompts as *random* semantically correct prompts. Such prompts are in

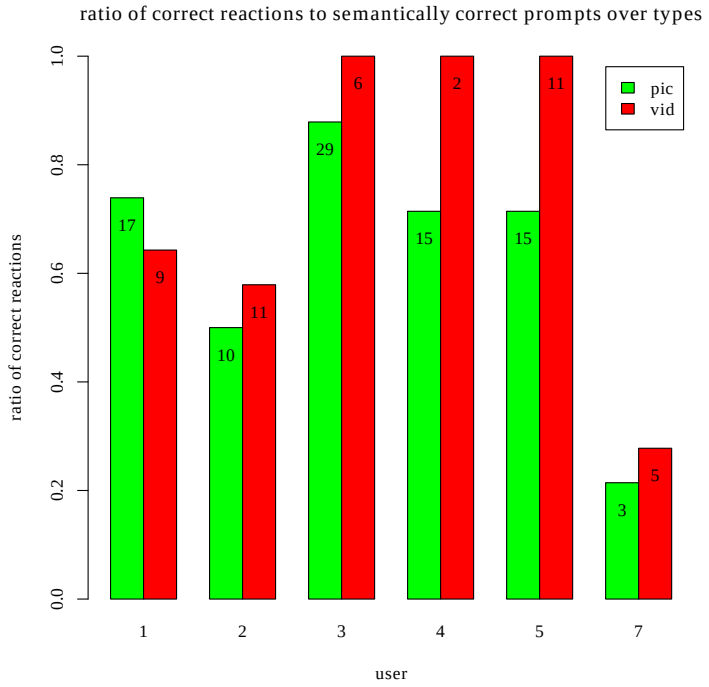


Fig. 16. Ratio of correct reactions to semantically correct pictogram and video prompts for individual users. pic - pictogram prompts, vid - video prompts.

contrast to *adequate* semantically correct prompts: adequate prompts help the user to initiate a next step when the user gets stuck in the task or interrupt an erroneous performance of the user during the task. Figure 16 contains both adequate and random semantically correct prompts. The figure contains a total number of 89 pictogram and 44 video prompts. User 3 shows a ratio of 89% correct reactions to pictogram prompts and 100% correct reactions to video prompts. Both kinds of prompts seem to be appropriate for user 3 with regard to the level of information provided in the prompts. Users 4 and 5 also show 100% correct reactions to video prompts, but only about 70% correct reactions to pictogram prompts. Pictogram prompts seem to be appropriate for users 4 and 5 in most situations. However, users 4 and 5 reacted incorrectly or not at all in 30% of the pictogram prompts. Video prompts seem to be more appropriate in such situations. Both user 4 and 5 reacted correctly in 100% of the cases. User 2 shows a different reaction behavior: the ratio of correct reactions to pictogram prompts is 50%. For video prompts, the ratio is increased with 60% correct reactions. Video prompts seem to be more appropriate compared to pictogram prompts. We found three possible explanations: firstly, video prompts are better suited to grab the attention of users than pictogram prompts because the movement in the videos is more salient than the static pictogram prompts. Some users might miss the static pictogram prompts. Secondly, users might be able to react to a video prompt due to priming effects: a user might already be primed by a pictogram prompt of the same behavior which timely precedes a video prompt in any case. Thirdly, a video prompt provides a higher level of information about the behavior. Hence, video prompts might be more suited to a user's cognitive abilities. We are not able to uncover the reasons from the results of

the study. We might investigate the reasons in more detail in future studies. The reaction behavior of user 7 is poor to both pictogram and video prompts with 20% and 25%, respectively. We found two possible explanations for the user's behavior: firstly, user 7 might not be able to react to prompts at all: both pictogram and video prompts seem to be inappropriate for user 7. Secondly, the user might not be willing to follow the prompts given by the TEBRA system. According to the caregivers, user 7 sticks to a strict routine in tooth brushing in which the user usually doesn't like distractions. This might indicate that user 7 is not willing to react to prompts. However, the exact reasons for the behavior of the user remain unclear. In the evaluation of inappropriate prompts, we didn't find any relationship between the number of inappropriate prompts and the number of reactions where a user ignores the system prompt.

The results show that the responsiveness to system prompts varies amongst individual users: some users react correctly to pictogram prompts, but other users need video prompts for proper assistance. The TEBRA system is able to deal with differences in the responsiveness of users by providing an escalation hierarchy which presents prompts with increasing level of information until the prompts provide appropriate assistance to a user.

A further aspect for the appropriateness of the TEBRA system is the evaluation of erroneous system behavior. We distinguish between two types of errors which lead to an erroneous system behavior: false-positives and false-negatives. False-positive errors (also called false alarms) happen when the system delivers a prompt, but the prompt is not necessary at that time. False-negative errors occur in situations where the system misses a prompt although a prompt would have been appropriate. Both types were manually annotated by the first author of the paper. False-positives were coded similarly to the annotation of the semantical correctness of prompts described earlier: when a prompt was issued, we compared whether the prompt was consistent with the overall progress of the user in the whole brushing task. For example, a prompt is inconsistent if the user has already performed the prompted behavior, but the system has not recognized it. False-negatives arise in situations where the user performs an inconsistent behavior, but the system didn't prompt the user.

Most of the erroneous prompts given by the system were prompts due to false-positive errors. We conclude that users accept false-positive errors when the system assists them properly throughout the remainder of the task by avoiding missing prompts (false-negatives). A trivial policy of avoiding false-negative prompts is providing prompts throughout the whole execution of the task. However, such a prompting behavior is not acceptable since the aim of an ATC system is increasing the independence of users by prompting when necessary. Hence, an appropriate prompting behavior requires a trade-off between minimizing false-negative prompts by providing steady prompting and increasing the independence of users by prompting when necessary. Future work might deal with this trade-off by studying different levels of 'prompting agility' of the TEBRA system.

5.2.3. Usability aspects. The application of an ATC system highly depends on the usability of such a system. Usability in the context of ATC refers to the ease of use with regard to the overall goal of proper task assistance. The users' opinions are important in order to judge the usability of the TEBRA system. After each SYS trial, we asked the user whether the system was helpful in task execution using a questionnaire. The question was asked by a caregiver who rated the answer on a 5-point Likert scale with 1 being no assistance at all and 5 denoting very good assistance. The average value of the TEBRA system's helpfulness is 4.1. The left plot in figure 17 shows the distribution of answers on the 5-point Likert scale. Hence, the TEBRA system is helpful in task execution from a user's subjective point of view despite a number of semantically incorrect

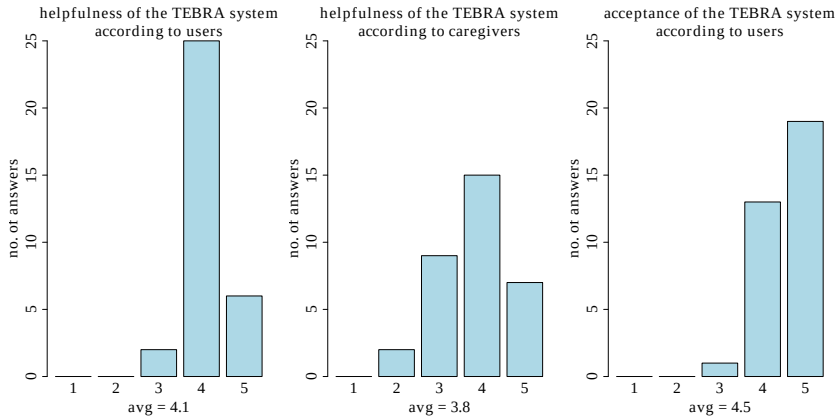


Fig. 17. Results of the questionnaire: helpfulness of the TEBRA system according to users (left plot) and caregivers (middle), acceptance of the TEBRA system according to users (right). Answers on a 5-point Likert scale with 1 being *not at all* and 5 denotes *very good*.

prompts due to perception errors. In addition to the users, we asked the caregivers to judge whether the system helped the user in the brushing task. The distribution of answers is shown in the middle plot of figure 17. The average value of 3.8 is lower compared to the user's opinion with 4.1. However, 3.8 is a good result which shows that the assistance of the TEBRA system is appropriate from an expert's point of view. A further aspect of usability is the user's acceptance of the system. We asked the users how much they liked to use the system as part of their daily routine. The right plot in figure 17 depicts the distribution of answers. An average value of 4.5 over all users underlines the good acceptance of the TEBRA system. Most of the users showed reactions such as smiling and laughing when they perceived system prompts. Furthermore, we observed that some of the users experienced the system as a kind of interaction partner: they talked to the system when a prompt was given or they reacted verbally to prompts by saying 'ok' or 'I will'. In two trials, we observed that the users were waiting with the execution of behaviors until the system prompted them what to do (due to a timeout). For these users, the interaction with the TEBRA system was a game-like situation where the users provoked a reaction of the system as an interaction partner. We encountered similar behavior in the CG trials where users tended to talk to the caregiver standing beside them. Users who talked frequently to the caregivers, were distracted more often and didn't focus on the proper execution of the task. According to the caregiver's comments, distraction due to verbal communication with the caregiver is one of the main sources for insufficient task execution. Since the TEBRA system is not able to respond to a user, the distraction due to verbal communication is minimized when using the TEBRA system. However, the communication between the caregiver and the user is an important social interaction for the user. Understanding the lack of such social interactions due to system use is an important issue in research of ATC systems, but is not taken into consideration in this article.

5.3. Threats to validity of the evaluation

The small sample size of six participants does not allow for hypotheses about the impact of the system for people with specific disabilities in general. The results presented are highly individual for different people. However, a trend with regard to a user's performance is clearly visible: the number of independent steps performed in the brushing task are increased for all participants. One could argue that the study design with

three CG trials in the first days and six SYS trials afterwards might have biased the task performance of users in a way that the participants have learned how to execute the task during the first trials. This is untenable due to the following reason: the group of participants in our study receive caregiver assistance in the brushing task throughout their whole life. According to the caregivers, very little or even no learning effect took place for those people in recent years. The very small number of three CG trials won't be able to bias a user's regular performance. Hence, we consider the CG trials as establishing a baseline, and we conclude that the effects in the SYS trials occurred due to the TEBRA system and not as a result of a learning effect.

The TEBRA system assists in tooth brushing which is one of many important tasks in a user's daily routine. Hence, the significance of the study results involve only the brushing task and might not generalize to other tasks such as dressing, shaving and cooking. However, due the modular implementation of the system, the TEBRA system is adjustable to assist in different tasks. The following steps would be necessary: firstly, an analysis of the task with Interaction Unit (IU) analysis needs to be conducted. From the results of the IU analysis, the initial design decisions regarding the sensor setup and the task execution framework for the planning component need to be extracted. The main components (behavior recognition and planning and decision making) need to be trained to the new task based on observational sample data. Additional studies with the TEBRA system assisting in different tasks would help to confirm the results presented here and to understand the general impact of ATC systems for people with cognitive disabilities.

6. CONCLUSION

This article has described the design, implementation and evaluation of the TEBRA (TEeth BRushing Assistance) system. TEBRA is a novel Assistive Technology for Cognition (ATC) for people with moderate cognitive disabilities. The TEBRA system provides assistance in the execution of brushing teeth by providing audio-visual prompts to users who are reliant on assistance in brushing teeth by a caregiver.

The main aim of the TEBRA system is to increase the independence of users from a human caregiver in the execution of brushing teeth. In order to evaluate its utility in this regard, we have conducted a study with seven people of the target group being assisted by a fully functioning prototype of the TEBRA system. The study data comprises 20 trials with a caregiver's assistance and 35 trials with the TEBRA system's assistance which is a large interaction corpus in the field of ATC. The results of the study showed that the TEBRA system is able to increase the independence of users in the tooth brushing task: all of the users were able to perform significantly more steps of the task independently, when they had been assisted by the TEBRA system instead of a human caregiver. The benefit of the system differs amongst users: one user showed only a slight increase of independent steps while another user was able to perform the brushing task completely independently in all trials with the system. However, also slight increases might be clinically meaningful for the users and their caregivers depending on the overall performance. The results of the study demonstrate the potential of the TEBRA system in assisting people with cognitive disabilities in task execution.

Future work includes two directions: firstly, the development of the TEBRA system towards a pervasive assistance system. Pervasive assistance refers to assistance in multiple tasks taking place at the washstand such as washing hands or shaving. An extension to multiple tasks raises further research problems: how will the TEBRA system be able to distinguish between different tasks rapidly in order to provide appropriate assistance from the very beginning of a task? How can the system cope with concurrent and interleaved execution of tasks? Secondly, the study results presented in this article are restricted to rather short-term effects in individual trials of users because

the study covered a period of five weeks. Long-term effects using the TEBRA system such as an increase in task performance for individual users over several months or years still need to be investigated in longitudinal studies in which a system is deployed for a longer period of time.

ACKNOWLEDGMENTS

The authors would like to thank the inhabitants and caregivers of Haus Bersaba for their high motivation to participate in the user study.

REFERENCES

M. F. Bevans and E. M. Sternberg. 2012. Caregiving burden, stress, and health effects among family caregivers of adult cancer patients. *American Medical Association* 307, 4 (2012), 398–403.

M.A. Demchak. 1990. Response Prompting and Fading Methods: A Review. *American Journal on Mental Retardation* 94, 6 (1990), 603–615.

A. Gillespie, C. Best, and B. O'Neill. 2011. Cognitive Function and Assistive Technology for Cognition: A Systematic Review. *International Neuropsychological Society* 18, 1 (2011), 1–19.

J.D. Gould and C. Lewis. 1985. Designing for usability: key principles and what designers think. *Commun. ACM* 28, 3 (1985), 300–311.

J. Hoey, T. Plötz, D. Jackson, A. Monk, C. Pham, and P. Olivier. 2011. Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing* 7, 3 (2011), 299–318.

J. Hoey, P. Poupart, A. von Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis. 2010. Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. *Computer Vision and Image Understanding* 114, 5 (2010), 503–519.

S. S. Intille, L. Bao, E. M. Tapia, and J. Rondoni. 2004. Acquiring in situ training data for context-aware ubiquitous computing applications.. In *CHI'04, Conf. on Human Factors in Computing Systems*. 1–8.

A. Kluger, J. G. Gianutsos, J. Golomb, S. H. Ferris, A. E. George, E. Franssen, and B. Reisberg. 1997. Patterns of Motor Impairment in Normal Aging, Mild Cognitive Decline, and Early Alzheimer' Disease. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* 52, 1 (1997), 28–39.

G. Leroy. 2011. *Designing User Studies in Informatics*. Springer, London, UK.

R. Levinson. 1997. The Planning and Execution Assistant and Trainer (PEAT). *Head Trauma Rehabilitation* 12, 2 (1997), 85–91.

E. F. LoPresti, A. Mihailidis, and N. Kirsch. 2004. Assistive technology for cognitive rehabilitation: State of the art. *Neuropsychological Rehabilitation* 14, 1-2 (2004), 5–39.

M. Melonis, Alex Mihailidis, Ryan Keyfitz, Marek Grześ, Jesse Hoey, and Cathy Bodine. 2012. Empowering adults with cognitive disability through inclusion of non-linear context aware prompting technology (N-CAPS). In *Proceedings of RESNA*.

A. Mihailidis, J. Boger, T. Craig, and J. Hoey. 2008. The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics (online)* 8, 28 (2008).

B. O'Neill and A. Gillespie. 2008. Simulating naturalistic instruction: the case for a voice mediated interface for assistive technology for cognition. *Assistive Technologies* 2, 2 (2008), 22–31.

B. O'Neill, K. Moran, and A. Gillespie. 2010. Scaffolding rehabilitation behaviour using a voice-mediated assistive technology for cognition. *Neuropsychological Rehabilitation* 20, 4 (2010), 509–527.

C. Peters, T. Hermann, and S. Wachsmuth. 2011. Prototyping of an automatic prompting system for a residential home. In *RESNA/ICTA 2011 Conf. Proc. (online)*.

C. Peters, T. Hermann, and S. Wachsmuth. 2012. User Behavior Recognition for an Automatic Prompting System - A Structured Approach based on Task Analysis. In *ICPRAM'12, Int. Conf. on Pattern Recognition Applications and Methods*. 162–171.

C. Peters, T. Hermann, and S. Wachsmuth. 2013. TEBRA - An automatic prompting system for persons with cognitive disabilities in brushing teeth.. In *HealthInf'13, Int. Conf. on Health Informatics*. 12–23.

M. E. Pollack, L. E. Brown, D. Colbry, C E. McCarthy, C. Orosz, B. Peintner, S. Ramakrishnan, and I. Tsamardinos. 2003. Autominder: an intelligent cognitive orthotic system for people with memory impairment. *Robotics and Autonomous Systems* 44, 3–4 (2003), 273–282.

S. B. Richards, R. L. Taylor, R. Ramasamy, and R.Y. Richards. 1998. *Single Subject Research and Designs: Applications in Educational and Clinical Settings*. Singular Publ. Group.

C. Robson. 2002. *Real world research: a resource for social scientists and practitioner-researchers*. John Wiley & Sons.

H. Ryu and A. Monk. 2009. Interaction Unit Analysis: A New Interaction Design Framework. *Human-Computer Interaction* 24, 4 (2009), 367–407.

M. J. Scherer, T. Hart, N. Kirsch, and M. Schulthesis. 2005. Assistive Technologies for Cognitive Disabilities. *Critical Reviews in Physical and Rehabilitation Medicine* 17, 3 (2005), 195–215.

A. M. Seelye, M. Schmitter-Edgecombe, B. Das, and D. J. Cook. 2012. Application of cognitive rehabilitation theory to the development of smart prompting technologies. *IEEE Reviews in Biomedical Engineering* 5 (2012), 29–44.

F. Siepmann, L. Ziegler, M. Kortkamp, and S. Wachsmuth. 2012. Deploying a modeling framework for reusable robot behavior to enable informed strategies for domestic service robots. *Robotics and Autonomous Systems (in press)* (2012).