Learning to recognise behaviours of persons with dementia using multiple cues in an HMM-based approach

Christian Peters, Sven Wachsmuth
Applied Informatics
Bielefeld University
33594 Bielefeld, Germany
cpeters,swachsmu@techfak.uni-bielefeld.de

Jesse Hoey
School of Computing
University of Dundee
DD14HN, Dundee, Scotland
jessehoey@computing.dundee.ac.uk

ABSTRACT

This paper presents a learning technique for visual event recognition in a system that assists persons with dementia during handwashing. The challenge is that persons with dementia present a wide variety of behaviors during a single task, typically changing their behaviours drastically from day to day. Any attempt at modeling this variety requires a large set of features, image regions, and temporal dynamics. In this paper, we approach this challenge by supervised learning of generative models from manually segmented and labelled video sequences. Our method uses a generic set of appearance-based colour, motion and texture features, over a static set of regions. We then present two HMM architectures that incorporate multiple image regions by either fusing on a feature-level, or later in the recognition process using a mixture-of-experts approach, in which a gating HMM is applied for the dynamic selection between specialised expert HMMs. Our models are trained on a clinical database of videos, and we compare the HMM approaches with a nearest neighbours scheme. Our results confirm the challenge we present, and indicate that our generative modelling techniques are suitable for inclusion in future prototypes of the hand washing assistant.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*; K.4.2 [**Computers and Society**]: Social Issues—*Assistive technologies for persons with disabilities*

General Terms

Algorithms,Human Factors

Keywords

Event recognition, Supervised learning, HMM, Task assistance

PETRA '09, June 09-13, 2009, Corfu, Greece
Copyright 2009 ACM ISBN 978-1-60558-409-6 ...$5.00.

1. INTRODUCTION

The average age of people in the industrialised countries will increase steadily in the next years and decades. Simultaneously, the number of persons suffering from cognitive disabilities such as *dementia* and *Alzheimer's* is on the increase. Persons with dementia often have difficulty accomplishing *activities of daily living* (ADLs) without the help of a caregiver. This leads to a loss of independence, and an (often intolerable) increase in burden on carers. Assistive technology can support people during ADLs, decreasing caregiver burden and increasing independence and quality of life for the sufferers [27]. Such technology monitors a person's activities, detects when assistance is needed, and provides help to guide ADL [25, 12, 19], to remind[15], and to share/transmit information and provide point-of-contact care [3]. Cameras are an important element for such technology, as a key requirement is *non-invasiveness*: persons with dementia cannot tolerate any wearable sensors or prescribed behaviors. While successful non-invasive assistive devices have been tested for specific tasks, their long-term economic and social acceptance across a wide range of tasks depends on the ability of the devices to *adapt* to users as needs, environments, and abilities change.

A key factor for adaptivity is the ability to learn behaviors of humans from data. Cognitive disabilities make this task very challenging, as behaviors of a single user can change significantly over time. Patterns of regular activity are broken: a person with moderate or severe dementia attempting to wash their hands, for example, may have little memory of having done so in the past. Therefore, their behaviors may exhibit a large variance in temporal and spatial dynamics which is depicted in figure 5 and 6. The image sequences show two *take soap* events and two *turn water off* events respectively which are performed by one person in each case. The sequences not only differ vastly in duration, but also in the execution style which is indicated by the movement of the hands. This variance causes significant problems for systems that do not learn, as it is very difficult to predict *a priori* what a particular user will do. This is in contrast to applications such as sign language recognition, where gestures are consistent and repeatable with (relatively) low variance. This paper takes a step towards learning human behavior categories from visual data, specifically for the case of an assistive device that helps persons with dementia complete the task of *hand washing* independently.

The device in question uses a single camera, placed overhead of a sink, and watches a person with dementia as they wash their hands. The system needs to recognize key behaviors of the user (e.g. *using the soap*), and use this recognition to estimate the stage the user is currently at. If the user stops making progress, the system can deliver an audio-visual cue, or prompt, to remind the user of what to do. These cues are often sufficient to re-engage the user [19]. Currently, the system uses a generic hand tracker [11], and a heuristic method for recognition of behaviors based on the temporal trajectories of hands [12]. The problem with this approach is that it does not generalise across users and from day to day, as persons with dementia can exhibit widely variant behaviors. For example, a person may one day turn on the water quickly and confidently, and the next go back and forth many times, adjusting the water again and again. Furthermore, this approach does not incorporate learning of behavior models.

This paper approaches the learning problem in a supervised approach, in which a dataset of six subjects washing their hands are temporally segmented and labelled with a pre-defined set of four behaviors: *using soap*, *turning water on/off*, and *rinsing hands.* A set of generic features are then extracted from each of a set of four static image regions. These features are then used as observations for a set of hidden Markov models (HMMs). We apply HMMs in order to represent the temporal structure of different behaviors in terms of feature observations. The HMMs are trained on part of the data, and then used to recognise behaviors in a test data set. We use a cross-validation approach to maximize the amount of test data, and test across all six subjects. We found the best peforming method to give an average recognition of 79% across all subjects and all behaviors.

The contributions of our paper are two-fold. First, we present a challenging and important learning problem: the behaviors of persons with dementia. We demonstrate two HMM-based approaches to this problem, and pinpoint the challenges involved in using such activity recognition in practical devices with a timely and relevant purpose. Second, we perform a comparison of approaches for combining data across multiple regions and multiple subjects, for data with a wide variation in dynamics. We do not, in this paper, address the question of invariance with respect to execution rate or camera viewpoint [28], leaving this for future work.

2. RELATED WORK

A functional taxonomy of human action recognition work identifies four categories: *initialization*, *tracking*, *pose estimation*, and *recognition* [20]. This paper presents work in the last category only. We are interested in recognising human behaviors with a particular *function*: to accomplish a sub-task in an activity of daily living (ADL). The functionality of our recognised behaviors is defined through their inclusion in a higher level model, as described in [12]. Other computer vision researchers have looked at activity recognition using top-down information (e.g. [14]), but we sidestep this issue here to focus only on recognition of predefined behaviors, while ignoring higher level information.

The recognition of predefined categories of visual activities has a strong tradition in computer vision [24, 20]. The work has spanned many types of behaviors for different applications in sports analysis, smart rooms, surveillance, and biometrics. Recent work by Efros *et al.* showed the importance of good features [4] for noisy action recognition, in a simple nearest-neighbors type recognition approach. We include these features in our set, and compare our HMM-based approaches with a similar type of nearest-neighbour recognition method. Other approaches use templates [5, 1], but are more applicable for well-defined gestures.

Hidden Markov model (HMM) or dynamic Bayesian network (DBN) based approaches have been popular. Yamato et al.[31] first used HMMs, which were later popularized in the recognition of American sign language [16, 29]. More recently, Oliver et al.[22] used a predefined hierarchical HMM to learn patterns of behavior in an office. Galata et al. use a variable length Markov model to recognize categories of dance movements [7]. Other variants of HMMs have been used for more complex scenes with interaction between regions and objects [2, 8]. DBNs have recently been used for characterization of human trajectories [21], and human interactions [23]. A multi-class support vector machine (SVM) classifier was used in[18] to classify the stages of lathering soap during handwashing of healthy subjects, with a view to quality assessment. This type of analysis is too detailed for our application, and would suffer due to large variabilities in behaviours in our target population. Few authors, however, have focussed on behaviors with large variances across users and time.

Learning of human behaviors from visual data has yet to be applied to assistive technologies. Instead, technologists have built simple, non-adaptive devices for specific tasks [12, 19, 25, 15]. Learning has been used to for sequential patterns of spatial locations in a home or outdoors [21, 17], or of precise usage activities [9]. The COACH is a real-time system that monitors a user with a video camera, provides audio-visual prompts to help a user in hand washing, and calls for human assistance if needed [12]. The device uses only video, combining computer vision for tracking hands with planning based on decision theory and probabilistic reasoning [12], but does not learn from data. It has been trialled with six subjects, and a dataset has been generated that is used in this paper [1].

A comprehensive survey of recent research into human action recognition [20] points to only few works that are attempting to qualify and handle the variations in human actions due to variations between individuals. Recent work has tried to quantify variations due to viewpoint, anthropometry, and execution rate [28, 30], but have not yet approached differences in human psychology. It is assumed that a cognitively able human is performing an action that many would perform in a similar way (e.g walking or running). Our case is different. We are considering humans who perform the same action (e.g. using the soap) in a very different way from day to day, or in relation to others. An example will help to clarify. A handwashing subject in a previous set of trials had a peculiar way of using the soap: with one hand, he would pump soap onto the counter-top. Then, after a long pause, he would retrieve the spilled soap with the same hand. We do not pretend to be able to recognise this radically different behavior, but use this example to underline the fact that the functional effects of a human behavior may be quite separate from the behavioral patterns, making recognition very challenging. This is an area in human action recognition in computer vision that has not yet been explored, and one that this paper attempts to expose.

[1] We acknowledge the the Univ. of Toronto for sharing this dataset.

3. APPROACH

Because of the wide variety of behaviors observed in persons with dementia, we neglect the trajectory information of the hands and focus on a fixed set of image regions that correspond to relevant parts in the scene. For each image region and each frame, a high-dimensional feature vector is computed that captures the color and intensity distribution, texture, and motion. Thus, the scene patches provide local cues that may be relevant or not for recognizing a specific event. The presented approach learns the dynamics as well as the relevance of each cue and combines them to a purely data-driven event recognizer. The dynamics are modeled by Hidden Markov Models (HMMs) where we explore different architectures. The overall approach makes the following assumptions. First, the image regions are pre-defined but they may slightly vary from trial to trial due to a segmentation by hand. All trials are perceived from a similar perspective. Although the automatic segmentation of events can be covered by the HMM approach, in principle, we concentrate on the classification of pre-segmented events in order to keep the comparability of alternative models.

We apply our approach in a handwashing scenario recognizing the following events: *take soap*, *turn water on/off* and *rinse hands*. An event called *no event* is included which describes the periods of time in which none of the other events happen. We consider four regions *soap*, *water*, *left tap* and *right tap*. Since the four regions are pre-segmented by hand for every trial video, the sizes and positions of these base regions don't form a perfect segmentation result, but differ slightly from trial to trial. In order to increase the robustness of the recognition, regions which are up- and downsized by a factor of 1.1 are also taken into account. Hence, we consider 12 regions per video frame which are composed of the four base regions plus the up- and downsized ones. We represent an event in terms of sequences of 16-dimensional feature vectors where each feature vector is calculated on each of the 12 regions for every frame of a trial video. We use generic features taken from the color/intensity, texture and motion domain as follows.

3.1 Color/Intensity features

Event recognition is highly associated to the detection of the activity of a person's hands. Since the hand's activity gives rise to a characteristic color- and intensity appearance the employment of color- and intensity features is reasonable. On a grayscale histogram we calculate four central moments: mean μ, variance V, skewness S determined by

$$S = \frac{1}{N \cdot \sigma^3} \sum_{i=1}^{M} (n_i - \mu)^3 \tag{1}$$

and kurtosis K which is

$$K = (\frac{1}{N \cdot \sigma^4} \sum_{i=1}^{M} (n_i - \mu)^4) - 3. \tag{2}$$

$\sigma = \sqrt{V}$ is the standard deviation, N the total number of pixels, n_i the number of pixels with grayscale i and $M = 256$ the number of grayscales.

We make use of the color information provided by calculating the difference of RGB color histograms taken at time t and $t-1$ respectively in order to capture significant changes in the color appearance. In the calculation of grayscale as well as color histograms, spatial information gets lost. Correlograms proposed by Huang et al.[13] combine this with intensity information. A correlogram is a three-dimensional matrix $C(i,j,d)$ where the entry (i,j,d) is the probability of finding a pixel with intensity i and a pixel with intensity j in a distance of d to each other. The difference of two correlograms calculated at time t and $t-1$ respectively is used as a feature representing both spatial and temporal intensity changes. In total, we have six color/intensity features.

3.2 Texture features

Most events cause textural structures on certain image regions, e.g. flowing water – as the result of event *turn water on* – leads to a structure in the region *water*. A common representation for the calculation of textural features is a gray level co-occurrence matrix $M_{\Delta x, \Delta y}(i,j)$ where the entry (i,j) is the number of pixel pairs where one pixel has intensity i and the other intensity j and at the same time the pixels have an offset of Δx and Δy. In our approach, we use the offset parameters $\Delta x = 0$ and $\Delta y = 1$. On the co-occurrence matrix, we calculate the following six *Haralick-features* as stated in [10]: Energy, entropy, contrast, homogeneity, inverse difference moment and correlation.

3.3 Motion features

Motion features are important since any event is initiated by the motion of a persons' hands. The motion features are based on the *spatio-temporal motion descriptors* proposed by Efros et al.[4]: An optical flow field F is calculated between two consecutive frames which is splitted into horizontal and vertical components F_x and F_y. The components are half-wave rectified into four non-negative channels F_x^+, F_x^-, F_y^+ and F_y^-. The difference f_{diff} of the four blurred motion channels Fb_x^+, Fb_x^-, Fb_y^+ and Fb_y^- computed at time t and $t-1$ are used as features resulting in four motion features.

3.4 Feature preprocessing

The features $f_1, \ldots, f_L$ used throughout this approach have both different co-domains and different variances. In order to avoid any bias in the following dimension reduction, a variance normalization is applied to each feature, individually. Thus, the normalized feature vector $\hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_L)$ is defined as

$$\hat{f}_i = \frac{f_i - \mu_i}{\sigma_i}, i = 1 \ldots L \tag{3}$$

where μ_i is the mean and σ_i the std. dev. of feature i. Finally, a Principal Component Analysis (PCA) is applied to the feature vectors in order to reduce feature correlations while keeping 90% of the energy.

In our approaches, we apply a linear HMM for each of the five events which are combined to a Compound HMM by disjunction, as depicted in figure 1. Each HMM consists of a sequence of states S_1 to $S_{N(ev)}$ where the length $N(ev)$ of the state sequence varies for each event type $ev \in \{\text{take\_soap}, \ldots\}$. The blank nodes denote the start and the end state of the parallelized HMMs. The HMMs are trained using the Baum-Welch algorithm which maximizes the production probability of an HMM given a set of sample events.

In this paper, the proposed features are used in two different HMM architectures that incorporate multiple image regions by either fusing them on a feature-level, or later in the recognition process using a mixture-of-experts approach for the dynamic selection between specialized expert HMMs.

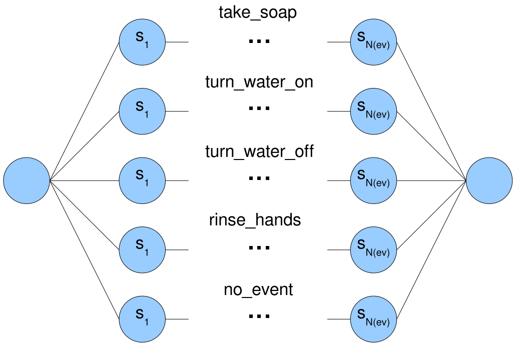

Figure 1: The compound HMM is a disjunction of linear HMMs.

3.5 Combination on feature level

In this approach, we directly learn the dependencies between events and regions on feature level by concatenating the 16-dimensional feature vectors calculated on the image regions *water*, *soap*, *left tap* and *right tap* to a 64-dimensional feature vector. A PCA reduces the feature vectors to 23 dimensions keeping 90% of the overall energy. We apply a single Compound HMM which generates a common event hypothesis for all regions. For a sequence of feature vectors denoting an event, the optimal production probability is computed for each event type. Then, the event hypothesis is given by the most likely event type. Figure 2 depicts the approach:

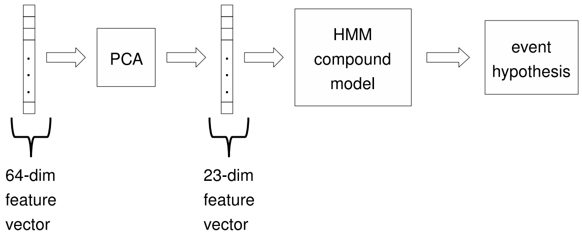

Figure 2: Combination on feature level using a single CHMM.

3.6 Mixture-of-experts combination

In this architecture, we use specialized experts for each of the four regions by applying one Compound HMM per region. Each Compound HMM is treated independently and generates a local event hypothesis based on the sequence of feature vectors calculated on the corresponding image region. The Compound HMM responsible for the *water* region, for example, is fed only by feature vectors calculated on the base, up- and downsized region *water*. For an overview of the mixture-of-experts method, see [26]. A PCA is applied to the feature vectors of each of the four regions, independently, resulting in 7-dim (water), 5-dim (soap), 7-dim (left tap), 8-dim (right tap) reduced feature vectors.

Most events affect multiple image regions, but will not cause characteristic scene changes in all of the four regions, simultaneously: *Take soap* will be detectable on the region *soap* but it will not affect any systematic changes on the *right tap* region. This information is not given to the system beforehand, but needs to be learnt. As a consequence, simple majority voting on the four regions will not work because recognition results on irrelevant regions will either be very brittle or just *no event*. In order to deal with this problem and to combine the local event hypotheses to a common hypothesis, we apply a mixture-of-experts approach: For each of the four regions, we train a second HMM that decides whether the local event hypothesis of the corresponding region is reliable and, therefore, regarded in the common event hypothesis. In the following, we refer to these HMMs as *activity HMMs* and to the Compound HMMs generating a local event hypothesis as *event HMMs*. In order to avoid brittle thresholding, an activity HMM (illustrated in figure 3) consists of a disjunction of two linear HMMs which are denoted *active* and *inactive*, respectively.

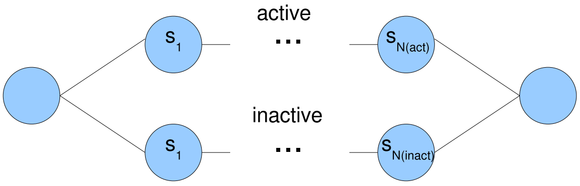

Figure 3: Activity CHMM consisting of two HMMs.

The training of these HMMs uses the global 23-dimensional feature vectors described in the previous section. First, the event HMMs are trained based on the local feature vectors. Then, these are used in order to separate the set of sample events into two disjoint sets based on the correct or incorrect classification result. The set on which the HMM *active* is trained is composed of correctly classified sample events. The training set for the *inactive* HMM contains events that were incorrectly classified. The training is shown for one region in figure 4. Hence, each activity HMM separates the space of 23-dimensional global input events, containing information from all regions, into the expectation of reliable and non-reliable local event hypotheses with regard to the associated region. In order to measure the degree of reliability, a weight w_R is calculated for every region $R \in \{\text{water, soap, left tap, right tap}\}$:

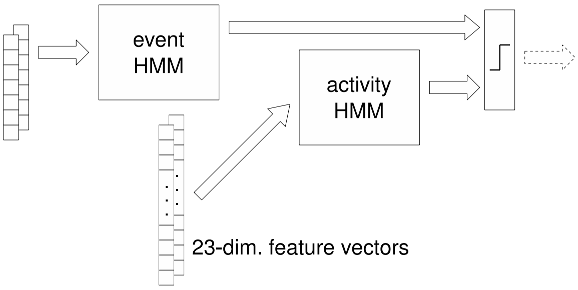

Figure 4: Event HMM trained based on local feature vectors, activity HMM trained based on global feature vectors.

$$w_R = 1 - \frac{s^R_{act}}{s^R_{act} + s^R_{inact}} \qquad (4)$$

where s^R_{act} and s^R_{inact} are the negative logarithms of the optimal production probabilities obtained by the corresponding activity HMM of region R. A hypothesis is considered in the common event hypothesis if the weight w_R is greater than 0.5: the optimal production probability for the *active* HMM is bigger than the optimal production probability for the *inactive* HMM. A combination of the hypotheses by majority vote leads to a common event hypothesis. If none of the event hypotheses pass the activity HMMs, the classification result is *no hypothesis* (NH) which describes the rejection of a sequence.

3.7 K-nearest neighbours

Each event is represented by a sequence of feature vectors where the length of the sequence is the number of frames of the trial video. In our approach, we calculate the three nearest neighbours of each feature vector according to the Euclidian distance, and combine them by majority vote for the event type to a hypothesis for each feature vector of the sequence. The recognised event type is then a majority vote over all feature vectors resulting in a common event hypothesis.

4. EXPERIMENTAL RESULTS

The experiments throughout this work are based on sample events taken from real clinical trials obtained by the existing handwashing system [19]. These trials were a modified withdrawal-type, single-subject test consisting of a baseline phase (no computer guidance), A, and an intervention phase (with computer guidance), B, tested in the order $A1 - B1 - A2 - B2$. Participants were recruited from a long-term care facility in Toronto, Canada, where the trials were conducted. The primary inclusion criteria were clinical diagnosis of moderate-to-severe dementia. Six older adults participated in the study—5 females, 1 male, average age 86.3 ± 8.8. Using the Mini-Mental State Examination [6], five of the subjects were classified as having moderate-level dementia, with the one remaining classified as severe. A fully functional washroom located in the long-term care unit was retrofitted with the necessary system hardware, specifically a ceiling-mounted IEEE-1394 digital video camera (Point Grey Research DragonFly2), and a Dell Latitude laptop computer (2 GHz processor, 2 Gb RAM). The video from the overhead camera was recorded on the laptop with a framerate of approximately 30Hz.

We initially selected 52 handwashing trials randomly. Figure 5 and 6 show image sequences taken from these trials. The subset of trials was then manually segmented in time into the five categories of events: using soap, turning water on, turning water off, rinsing hands, and no event. This segmentation was performed by the first author in the following way: For every trial, the four regions *water*, *soap*, *left tap* and *right tap* were marked by a rectangle. Each event was then considered to start as the hands entered the main region for that event, and end when the hands left the region. Breaks between events, as well as events that can not be assigned to static regions, including lathering and drying the hands, are labeled as *no event*. The resulting dataset consisted of 589 events from the six subjects, with a distribution across the five event categories as shown in Table 1. The number of *no event* events is very high, since in a handwashing trials, *no event* usually alternates with one of the other events and therefore occurs more frequently because the subjects in the trials make frequent pauses during their execution. This is a typical style of behavior for persons with dementia. Furthermore, the varying behavior of persons with dementia can be shown in table 1 since events that normally occur once in handwashing trial occur several times per subject, for example take soap in subject 1.

Sub	trials	TS	WON	WOFF	RH	NE	Sum
1	6	8	8	8	6	36	66
3	14	14	19	20	28	88	172
4	10	8	8	10	11	42	79
5	5	8	6	6	8	33	61
6	10	10	10	10	36	85	121
8	7	6	10	9	18	47	90
Sum	52	57	61	63	97	311	589

Table 1: Number of events of the different subjects. The events are labelled take soap (TS), turn water on (WON), turn water off (WOFF), rinse hands (RH) and no event (NE).

	length in frames			**number**	
event	min.	max.	median	*all*	*most*
take soap	40	70	54	57	30
turn water on	20	40	30	61	20
turn water off	10	40	27	63	29
rinse hands	1	60	36	97	23
no event	1	100	11	311	177
overall	1	100	27	589	279

Table 2: Overview of the length ranges per event in which most of the events reside and the number of different events.

As the variance in event lengths will impact the HMM classification, we constructed a subset of the sequences in which the lengths are constrained to a range in which most of the events reside based on the length histograms. Table 2 shows the minimum, maximum and median lengths as well as the number of events used in this constrained dataset. We refer to the complete set of 589 events as the *all* dataset, and the constrained dataset as the *most* dataset.

We conducted three types of experiments, in each of which we applied leave-one-out cross-validation in order to construct the classification results from a sufficient number of test events calculated on base regions. The three experiments are as follows

- **all-subjects** (**AS**): all subjects' sequences were used together. A single sequence was removed in the cross-validation, and the models were trained on all remaining sequences.
- **single-subject** (**SS**): data from a single subject were used for training and testing. Table 1 gives an overview of the number of events per subject.
- **leave-one-subject-out** (**LO**) in which the training data was taken from 5 subjects, and the test data from the sixth. In this case, only a single experiment was performed per subject (6 experiments in all).

Finally, we conduct experiments using the three methods described above: **FC** refers to the feature combination method (Section 3.5, **ME** refers to the mixture of experts method (Section 3.6) and **KN** refers to the k-nearest-neighbors approach (Section 3.7).

Table 3 shows average classification rates for all three experiments, both data sets *all* and *most*, for the three methods **FC,ME,KN**. The results of the *most* dataset in conjunction with the SS experiment are dropped since the amount of training and test samples is too small.

exp	data	meth.	TS	WON	WOFF	RH	NE (NH)
AS	all	FC	82.5	**56.7**	**58.7**	**96.9**	80.7
		ME	**86**	49.2	38.1	72.2	66.2 (15.8)
		KN	61.4	55.7	54	83.5	**88.1**
	most	FC	80	63.2	**72.4**	**87**	**93.2**
		ME	**83.3**	60	65.5	78.3	89.3 (0.6)
		KN	66.7	**75**	62	60.9	88.1
SS	all	FC	**68**	48	**56.6**	66.3	**86.1**
		ME	59.3	45.6	55.6	55.1	66 (14.7)
		KN	64.3	**48.8**	44	**69.9**	76.5
LO	all	FC	**82.5**	33.3	**63.5**	**93.8**	77.2
		ME	**82.5**	**50.8**	49.2	84.5	61.7 (12.5)
		KN	62.3	37.1	53.3	72.2	**83.6**
	most	FC	**76.7**	31.6	48.3	**87**	**92.1**
		ME	60	15	37.9	69.6	88.7 (0.6)
		KN	41	**53**	**70.4**	52	81.7

Table 3: Classification rates of the two expriments AS: all subjects, SS: single subject, LO: leave-one-subject-out, for the two data sets *all* and *most*, and the three methods: feature combination (FC), mixture-of-experts (ME) and k-nearest-neighbours (KN) approaches. The events are labelled take soap (TS), turn water on (WON), turn water off (WOFF), rinse hands (RH), no event (NE). The number in brackets after no event is the ratio of no event classified as no hypothesis (NH).

In general, the experiment using all subjects together (AS) leads to better classification rates compared to the single-subject (SS) and the leave-one-subject-out (LO) experiments. In the SS experiment, the number of training items is not sufficient to learn specific models for each person which leads to worse results compared to the AS experiment. The classification rates for *all* and *most* datasets are similar for the events *take soap* and *rinse hands*. However, the rates for the events *turn water on* and *turn water off* show a significant advantage for *most* experiment. For these events, timing is an important factor since extremely short and long events cause an increased error rate. We also see that the FC approach leads to better classification results than the ME and the KN approach with regard to a reduced training set in both SS and LO approaches.

For a more detailed comparison of the average classification rates of the feature combination (FC) and the mixture-of-experts (ME) approach with the *all* dataset, we take the confusion matrices into account which are depicted in tables 4 and 5. In the following, the events are labelled take soap (TS), turn water on (WON), turn water off (WOFF), no event (NE), no hypothesis (NH). Although the average classification rates are increased for the FC compared to the ME approach, the false-positive rates are decreased in the ME approach. Especially for the events *turn water on* and *turn water off*, the false positives are decreased from 38.3 to 14.8, and from 38.1 to 19 respectively. Instead of misclassifying, the ME approach more often rejects events as no hypothesis (no hyp). Especially in a prompting system that assists users in a specific task, the minimization of misclassifications as well as a rejection of vague hypothesis are desirable.

	TS	WON	WOFF	RH	NE
TS	82.5%	5.3%	0%	0%	12.3%
WON	0%	56.7%	38.3%	3.3%	1.7%
WOFF	0%	38.1%	58.7%	1.6%	1.6%
RH	0%	0%	1%	96.9%	2.1%
NE	4.2%	0.6%	4.2%	10.3%	80.7%

Table 4: Confusion matrix of the feature combination approach using the *all* dataset in the AS experiment.

	TS	WON	WOFF	RH	NE	NH
TS	86%	1.8%	0%	0%	5.3%	7%
WON	0%	49.2%	14.8%	1.6%	4.9%	29.5%
WOFF	0%	19%	38.1%	0%	17.5%	25.4%
RH	1%	1%	4.1%	72.2%	6.2%	15.5%
NE	3.9%	1%	3.5%	9.6%	66.2%	15.8%

Table 5: Confusion matrix of the mixture-of-experts approach using the *all* dataset in the AS experiment.

The average results for the single-subject experiments shown in Table 3 hide significant variation amongst subjects. Table 6 shows the classification rates of the both the feature combination (FC) and mixture-of-experts (ME) approach listed for individual subjects. The total classification rates are composed of extremely varying classification rates for the different subjects. Subject 1 has very poor classification rates since this subject is highly dependent on the help of

meth.	Sub	TS	WON	WOFF	RH	NE	NH
FC	1	25	28.6	50	66.7	75	
	3	88.2	27.8	65	100	80.7	
	4	87.5	85.7	30	100	83.3	
	5	87.5	40	50	100	84.8	
	6	100	88.9	80	100	70.8	
	8	100	88.9	66.7	94.4	93.6	
ME	1	75	28.6	12.5	0	52.8	27.8
	3	76.5	33.3	35	78.6	72.7	12.5
	4	75	28.6	30	63.6	66.7	9.5
	5	100	80	50	37.5	12.5	39.4
	6	100	88.9	80	80.8	66.2	13.8
	8	100	66.7	22.2	94.4	80.9	4.3

Table 6: Classification rates of the methods feature combination (FC) and mixture-of-experts (ME) listed for any subject with the AS experiment. (NH) is a possible output only for the ME approach since the FC approach does not incorporate a rejection.

a caregiver. Without this help, subject 1 was nearly unable to perform any handwashing event. Hence, the active help of a caregiver corrupts the feature representation of events which leads to poor classification rates. However, the results of poor classified subjects underline the challenge of behavior recognition in a scenario with persons suffering from dementia who show significant variabilities in both execution and duration of events as depicted in the image sequences in figures 5 and 6. On the other hand, subjects 6 and 8 produce excellent classification rates. The confusion matrix for subject 6 is depicted in table 7. Especially, the results of the events *turn water on* and *turn water off* are excellent compared to the existing trajectory-based approach of the handwashing system in which *turn water on* and *turn water off* can not be distinguished. The very good classification rates of single subjects show that our approaches perform well in scenarios in which the behaviors are more regular. It is important to underline that no top-down knowledge about the execution of the underlying task is included in our approaches. The results strengthen the hypothesis that the generic features in combination with the learning approaches can discriminate well between different behaviors.

	TS	WON	WOFF	RH	NE	NH
TS	100%	0%	0%	0%	0%	0%
WON	0%	88.9%	11.1%	0%	0%	0%
WOFF	0%	10%	80%	0%	10%	0%
RH	0%	0%	0%	80.8%	3.8%	15.4%
NE	1.5%	0%	1.5%	16.9%	66.2%	13.8%

Table 7: Confusion matrix for subject 6 in the mixture-of-experts approach using the *all* dataset.

5. CONCLUSION

In this paper, we target the challenging problem of recognizing task relevant events of highly variant performances of ADLs, specifically by persons with dementia. We learn our models in a supervised fashion using real video data from clinical trials in the context of an assistive handwasching scenario. We show that a purely data-driven recognition of task-relevant events is feasible based on features that are computed on multiple image regions. Therefore, we suggest two different HMM-based approaches and show them to be superior to a simple k-NN recognition scheme in most experiments. Our analysis shows that the recognition performance varies extremely between different subjects, stressing the overall challenge of these kind of datasets. The excellent results on some subjects show the potential of the HMM-based scheme to distinguish events like *turn-water-on* and *turn-water-off* which have a very similar appearance. These would not be distinguishable by approaches purely based on hand trajectories. Furthermore, the mixture-of-expert (ME) HMM approach reduces the number of false positives for more difficult subjects. Finally, both HMM architectures, but especially the ME HMM is shown to better generalize on very small training sets. Further work will concentrate on the combination with complementary approaches based on hand trajectories and the integration into an assistive system.

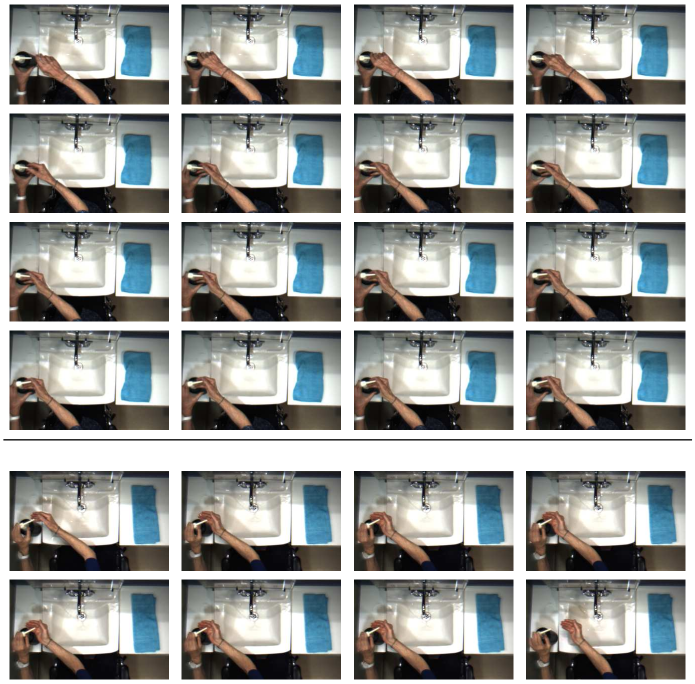

Figure 5: Two different take soap events from subject 5 having a length of 114 and 59 frames, respectively. Every $7th$ sequence frame is shown.

6. REFERENCES

[1] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, Mar. 2001.

[2] M. Brand. Coupled hidden markov models for modeling interacting processes. Technical report, 1996.

[3] P. Deegan, R. Grupen, A. Hanson, E. Horrell, S. Ou, E. Riseman, S. Sen, B. Thibodeau, A. Williams, and D. Xie. Mobile manipulators for assisted living in residential settings. *Autonomous Robots*, 24(2):179–192, February 2008.

[4] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE Intl. Conf. on Computer Vision*, pages 726–733, Nice, France, 2003.

[5] A. Elgammal, V. Shet, Y. Yacoob, and L. Davis. Learning dynamics for exemplar-based gesture recognition. In *Proc. CVPR*, Madison, WI, 2003.

[6] M. F. Folstein, S. E. Folstein, and P. R. McHugh. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, November 1975.

[7] A. Galata, N. Johnson, and D. Hogg. Learning variable-length Markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.

[8] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 742, 2003.

[9] V. Guralnik and K. Z. Haigh. Learning models of human behaviour with sequential patterns. In

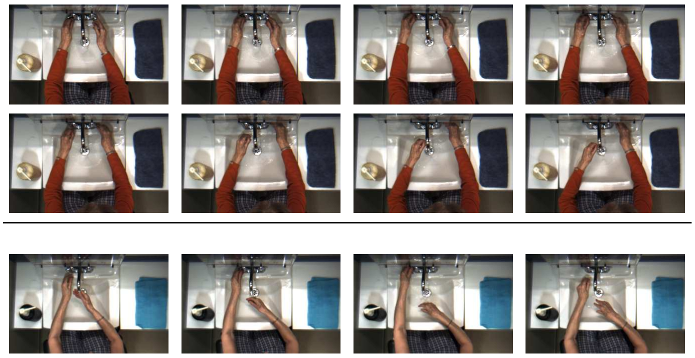

Figure 6: Two different turn water off events from subject 3 having a length of 57 and 24 frames, respectively. Every *7th* sequence frame is shown.

AAAI-02 Workshop on Automation as Caregiver, pages 24–30, July 2002.

[10] R. Haralick. Statistical and structural approaches to texture. 67(5):786–804, May 1979.

[11] J. Hoey. Tracking using flocks of features, with application to assisted handwashing. In M. J. Chantler, E. Trucco, and R. B. Fisher, editors, *Proc. of the British Machine Vision Conf.*, volume 1, pages 367–376, Edinburgh, Sept. 2006.

[12] J. Hoey, A. von Bertoldi, P. Poupart, and A. Mihailidis. Assisting persons with dementia duing handwashing using a partially observable markov decision process. In *Proc. of the Intl. Conf. on Vision Systems (ICVS)*, Bielefeld, Germany, 2007.

[13] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms, 1997.

[14] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[15] H. Kautz, L. Arnstein, G. Borriello, O. Etzioni, and D. Fox. An overview of the assisted cognition project. In *AAAI-2002 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, Edmonton, 2002.

[16] C. Lee and Y. Xu. Online, interactive learning of gestures for human/robot interfaces. In *In IEEE Intl. Conf. on Robotics and Automation*, pages 2982–2987, 1996.

[17] L. Liao, D. Fox, and H. Kautz. Learning and inferring transportation routines. In *Proc Nineteenth National Conf. on Artificial Intelligence (AAAI '04)*, pages 348–353, San Jose, CA, 2004.

[18] D. F. Llorca, F. Vilarino, J. Zhou, and G. Lacey. A multi-class svm classifier for automatic hand washing quality assessment. In *Proc. of the British Machine Vision Conference*, 2007.

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

[20] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, November-December 2006.

[21] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. H. Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In *Proc. CVPR*, San Diego, CA, June 2005.

[22] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Proc. of Intl. Conf. on Multimodal Interfaces*, Pittsburgh, PA, Oct. 2002.

[23] S. Park and J. Aggarwal. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, 10(2):164–179, 2004.

[24] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, 19(7):677–695, July 1997.

[25] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, 42(3–4), 2003.

[26] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[27] M. E. Pollack. Intelligent technology for an aging population: The use of ai to assist elders with cognitive impairment. *AI Magazine*, 26(2):9–24, Summer 2005.

[28] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *Proc. of Intl. Conf. on Computer Vision (ICCV)*, volume 1, pages 144–149, October 2005.

[29] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *In Intl. Workshop on Automatic Face and Gesture Recognition*, pages 189–194, 1995.

[30] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[31] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *Computer Vision and Pattern Recognition, 1992.*, pages 379–385, 1992.