Recognizing Daily Activities in Realistic Environments through Depth-Based User Tracking and Hidden Conditional Random Fields for MCI/AD Support

Dimitris Giakoumis[1], Georgios Stavropoulos[1,2], Dimitrios Kikidis[1], Manolis Vasileiadis[1], Konstantinos Votis[1], and Dimitrios Tzovaras[1]

[1] Information Technologies Institute, CERTH, Thessaloniki, Greece
[2] University of Patras, Patras, Greece

Abstract. This paper presents a novel framework for the automatic recognition of Activities of Daily Living (ADLs), such as cooking, eating, dishwashing and watching TV, based on depth video processing and Hidden Conditional Random Fields (HCRFs). Depth video is provided by low-cost RGB-D sensors unobtrusively installed in the house. The user's location, posture, as well as point cloud -based features related to gestures are extracted; a standing/sitting posture detector, as well as novel features expressing head and hand gestures are introduced herein. To model the target activities, we employed discriminative HCRFs and compared them to HMMs. Through experimental evaluation, HCRFs outperformed HMMs in location trajectories-based ADL detection. By fusing trajectories data with posture and the proposed gesture features, ADL detection performance was found to further improve, leading to recognition rates at the level of 90.5% for five target activities in a naturalistic home environment.

Keywords: ADL recognition, user location trajectories, posture, gestures, point-cloud features, hidden conditional random fields

1 Introduction

Automatic domestic activity recognition is a significant challenge, toward future homes equipped with robotic applications capable to monitor the resident's behaviour, identify abnormalities and assist in the establishment of daily activities [8]. This is of particular importance for cases of Mild Cognitive Impairments (MCI) or Alzheimer's Disease (AD), whereas activity monitoring can facilitate early diagnosis of cognitive decline [2]. Typically, the recognition of Activities of Daily Living (ADLs) [12] such as cooking, eating, dishwashing, has been approached through ambient sensors [3] monitoring the house environment [23], as well as locations visited from the monitored person [13][9]. During the last years, relevant research efforts have focused on RGB video processing [6][5][18][31] or, especially after the emergence of the Kinect sensor, on RGB-D images [30][4].

In practical applications, where the need for robust, continuous user tracking and reduced obtrusiveness is of major importance, low-cost depth sensors (e.g. Kinect) can play a vital role; they can provide the basis for simple, low-cost sensor networks capable to track the user's silhouette throughout the house. Such networks can be rather unobtrusive in terms of input data (privacy-preserving depth images), as well as in terms of installation set-up, since limited amount of cameras (e.g. one per room) can be installed for e.g. at room roof-top corners. Although computer vision for assistive robotic applications typically considers input taken from a (depth) camera installed on the robot [29], such approaches require the robot to continuously follow the user, maintaining an appropriate view angle that allows user actions tracking. Although effective to some extent, such approaches are rather difficult to guarantee continuous user tracking in realistic scenarios, where the monitored person moves freely throughout the house. Thus, strategically and unobtrusively installed low-cost depth sensors could provide auxiliary input to the robotic system, so as to establish more detailed, continuous user tracking.

Of course, reduced cost and installation complexity of such depth sensor networks comes at a cost; that of highly varying user viewing angle, noisy user point-cloud data and occlusions that pose significant challenges in practical user posture and gesture recognition for ADL detection. Since however user location can still be robustly tracked through silhouette extraction from depth images, past approaches for user location-based ADL detection, based however on complex multi-sensorial networks [9] or RGB video [6] should be revisited, examining the potential of their rationale to advance automatic ADL detection in this new sensor context. Herein, unobtrusive continuous user location tracking is available and can be fused with robust descriptors of user pose and gestures, capable to operate under the limitations of realistic settings, toward advancing the effectiveness of future, practically applicable ADL recognition systems.

The present work follows exactly this line, introducing an ADL recognition framework based on user location trajectories and moreover, on posture and novel descriptors of the user's point cloud, capturing characteristics of head and hand gestures. The proposed framework operates on the basis of Hidden Conditional Random Fields (HCRFs), building upon the capacity of such discriminative models to provide better recognition performance, compared to their generative counterparts, i.e. Hidden Markov Model (HMM) -based models [9][24].

1.1 Related Work

As different activities typically involve different house regions that are visited by the user, with different per-activity visit frequencies, several research works have demonstrated the feasibility of detecting ADLs through the 2D trajectories generated in the house floor plan as the user moves around. Indicatively, [6] used a two-layer Hidden Semi-Markov Model, to infer ADLs such as cooking or ironing, solely based on user location trajectories taken from RGB cameras. Moving away from computer vision, in [9], state-change sensors were used, indirectly indicating among others user location, providing input to either HMMs or CRFs

so as to discriminate among ADLs such as brushing teeth or cooking - preparing dinner. In [11], Semi-Markov HMM and CRF models were examined, again on the basis of state-change sensors. The work of [7], used again ambient sensors to recognize ADLs through a SVM-based detector. Although of potential toward effective ADL recognition, the above approaches, apart from [6], were based on complex multi-sensor systems. Thus, they need a large number of sensors to be installed in the house, whereas approaches based on computer vision [6], once their obtrusiveness is constrained and become privacy preserving, offer a major appealing characteristic; the fact that user activities can be tracked through only for e.g. one or two vision sensors installed in each monitored space.

Focusing on computer vision, toward incorporating apart from user location, information extracted from the user's silhouette, the work of [31] built on RGB video taken from a fisheye camera to detect activities following a three-level detection approach; the first level regarded user location and speed, the second body shape information for estimating the level of body motion and the third used primitive visual features to approach action recognition. A kNN-based detector was used to recognize cooking, brushing teeth and exercise activities. Building again upon RGB video, a system of two wide-field-of-view cameras and two narrow-field-of-view ones was employed in [18] to capture both coarse-level and fine-level activities respectively, utilizing a hierarchical Dynamic Bayesian Network (DBN). Although of potential, this line of approaches builds upon obtrusive RGB surveillance video processing.

Building upon posture and gestures, works such as [30][4] have explored RGB and depth video information fusion toward the recognition of daily actions, such as drink, pick up or sit down. In these works, features extracted from the depth video -based user's point cloud were proposed, toward 3D action representation; however, the RGB channel is still utilized, yielding privacy issues. Another line of research builds upon the depth-based markerless skeletal joints estimation method of [21]; in [28], skeleton-based body movement features were proposed for ADL recognition. Nevertheless, Kinect-based markerless skeletal joints extraction still suffer from varying view angles, occlusions and clutter in practical surveillance settings [30].

Focusing on the sole use of depth features for the detection of activities, a diverse set of approaches have been adopted by the scientific community, trying to increase the quality of the produced results while addressing specific experimental limitations. In a recent literature review on human activity recognition from 3D data [1], Aggarwall and Xia adopted a taxonomy of five types of features extracted from depth images, namely: 3D silhouettes, skeletal joints/body parts, local occupancy patterns, local spatio-temporal features and 3D optical flows. All these categories have shown promising results for the recognition of human activities but they were also found to be greatly influenced by difficulties found in real-life situations. Noise, object occlusion and camera position are some of the parameters that can dramatically degrade the results of the first three categories of features, whereas 3D optical flows and spatio-temporal features need colour information for reliable results.

Methods such as [16] have been found effective in the discrimination among different actions on rather controlled datasets (e.g. MSR Action 3D dataset [25]). However, by relying on the extraction of surface normals from the user's point cloud, they can easily become problematic in practical settings, where the user's point cloud can be highly noisy. In the present work, we focus on the recognition of daily activities in practical settings, where depth measurements are taken from un-optimal angles compared to datasets like [25] and are typically prone to occlusions and noise. In such settings, novel, more robust descriptors, capable to provide useful information to the ADL recognition system even in cases of significantly noisy user's point cloud are a significant challenge.

In the past, diverse classifiers have been used for recognizing daily activities, such as Bayesian networks [18] or SVMs [7]. As activity recognition is intrinsically a temporal classification problem [24], emphasis have been paid on Markovian state sequence models, such as HMMs [9] and their extensions employing either explicit state duration modelling [6] or hierarchical structures [10]. Relatively limited works have examined CRFs for ADL recognition [9], which can be seen as the discriminative counterparts of HMMs [24].

Conditional Random Fields are discriminative models for labelling sequences of observations. They condition on the entire observation sequence, while the features used as input can violate independence assumptions between observations, contrary to HMMs [24]. CRFs have been extensively used in the past toward gestures recognition [27][19] and motion tracking [22]. They have also been found effective in ADL detection based on state-change ambient sensors [9]. While CRFs generate per-observation labels, Hidden CRFs (HCRFs) [26] incorporate hidden states to model the underlying structure of the observations, providing a single label for the whole observation sequence. HCRFs have been successfully used in the past for gesture recognition [26], while by definition, they provide a potentially useful alternative to HMM-based approaches toward ADL recognition. Although the study of [24] indicated the potential of CRFs to drive location trajectories-based activity recognition by outperforming HMMs, to the best of our knowledge, such discriminative models and especially HCRFs, have still not been examined in the context of practical in-house ADL detection that builds upon user location trajectories.

1.2 Contribution

The present study follows the line of [9], where discriminative models, i.e. CRFs were found to have the potential to outperform generative HMMs in recognizing home ADLs through ambient state-change sensors. Considering that more fine-grained information regarding the user location, i.e. user location 2D trajectories on the house floor plan, has been found capable to drive ADL recognition through generative HMM-based models [6], this work first examines the capability of discriminative models, in particular Hidden CRFs (HCRF) [26], to advance effectiveness of ADL recognition, on the basis of user location trajectories that can be extracted from a small set of low-cost depth sensors installed unobtrusively in the house. Moreover, extending this line of research, we also incorporate

user posture information in the recognition scheme, as well as novel 3D point-cloud features of the user's silhouette that are herein introduced to express head and hand gestures. Through experimental evaluation with data derived from realistic house settings, HCRFs were found to outperform HMMs in detecting the target activities from user location trajectories only. By fusing trajectory-based features with user posture and our proposed gestural features, ADL recognition performance was found to further increase, reaching precision and recall at the level of 90.52% for five target activities.

1.3 Paper Outline

Section 2 presents our proposed method for detecting standing/sitting postures and the descriptors of the user's point-cloud, which capture information related to head and hand gestures. Section 3 describes our HCRF-based activity recognition framework and Section 4 describes the process that was followed for experimental evaluation and its findings. Conclusions are drawn in Section 5.

2 Depth Video-Based User Monitoring

In order to track user movement around the house and her/his posture and actions, one must first extract the user's silhouette from the depth input images. To this end, a background image is captured prior to our system's initialization, with the monitored area empty of moving objects. For each captured frame i during runtime, the binary user silhouette $S_i, i = 1...n$ is extracted, by subtracting the depth value of each pixel (x, y) of the background image BG, from its corresponding pixel in the current frame I_i. A pixel (x, y) is considered as foreground (silhouette) if its depth value differs from the corresponding background pixel's value by more than a predefined threshold T:

$$S_i(x,y) = \begin{cases} 0 \ \textit{(Background)} & \text{if } abs(I_i(x,y) - BG(x,y)) \leq T \\ 1 \ \textit{(Foreground)} & \text{otherwise} \end{cases} \quad (1)$$

After the silhouette is extracted, noise induced by small changes in the background, for e.g. movement of objects like chairs or tables, is removed through post processing. This is achieved by performing connected component analysis [20] on the foreground image and taking into account the position of the user in previous frames. This way, only the area containing the user is kept.

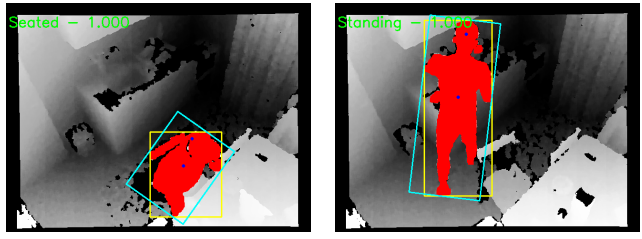
2.1 User Location and Posture Tracking

User location with respect to the house floor plan is trivially estimated on the basis of the relative position of the silhouette and the camera, as well as the known in-house camera position. In order to estimate whether the user is standing or sitting, an approach similar to [15], albeit more robust to occlusions, is followed. Using the silhouette image as a mask, the 3D point cloud of the user is extracted and transformed from the camera coordinate system, to the coordinate system

of the user, using the calibration information of the camera. Then, depending on the 3D point cloud's bounding box ratio $r = width/height$ and height h, the user's posture P_i for the frame i, is determined using a set of experimentally defined thresholds $h1, h2$(with $h1 > h2$) and $r1, r2$(with $r1 > r2$) for h and r respectively:

$$P_i = \begin{cases} Standing & \text{if } r_i < r_2 \ \ \& \ \ h_i > h_1 \\ Seated & \text{if } r_1 > r_i > r_2 \ \ \& \ \ h_1 > h_i > h_2 \end{cases} \quad (2)$$

Contrary to [15], the bounding box height is calculated in the proposed approach as the distance between the upper part of the silhouette from the floor, in the z axis of the building coordinate system. This allows robust estimation of the silhouette height, even in cases where the silhouette is occluded by objects lying between the user and the camera.



(a) Seated user (b) Standing user

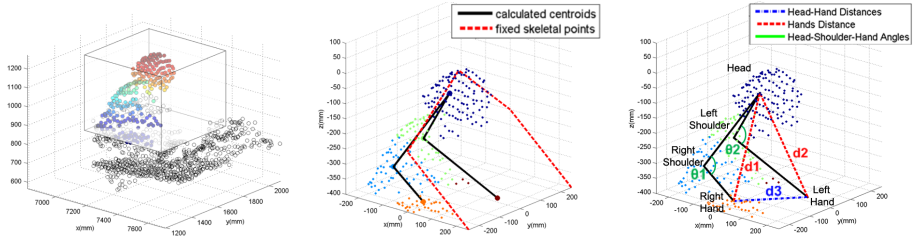
Fig. 1. Example of posture detection algorithm output

2.2 Point-Cloud Features of Upper-body Geometry and Gestures

The present study aims to address the difficulties of activity detection algorithms when used in uncontrolled, real life environments. To this end, a robust and privacy preserving algorithm is proposed that can provide input for recognizing high-level activities, focusing on point-cloud descriptors capable to encapsulate information related to the users posture and motion. Specifically, a set of six features is defined, with special focus on the recognition of eating activity, given its importance in MCI and AD patients; AD patients at later stages can have difficulties to successfully establish eating, due to short-term memory problems.

The proposed method aims at extracting depth features that can approximate the geometry of the user's upper body but can also be minimally influenced by occlusions, changes in the orientation of the user and capable to produce reliable results independently of the camera position. The framework for the calculation of these features is summarized in Fig. 2. First, the periods when the user is seated are kept and for each frame, a bounding box is defined to include the majority of points of the user's upper body and at the same time minimize the

influence of objects in the close proximity. Specifically, the horizontal dimensions of the bounding box are defined to be equal to the 2/3 of the user's arm length whereas its vertical dimension is taken as the 1/4 of the user's height. The position of the box is dynamically defined in every frame so that the centre of its upper side coincides with the highest point of the user's cloud (Fig. 2(a)).



(a) Bounding box (b) Point cloud clustering (c) Definition of features

Fig. 2. Framework for the extraction of features

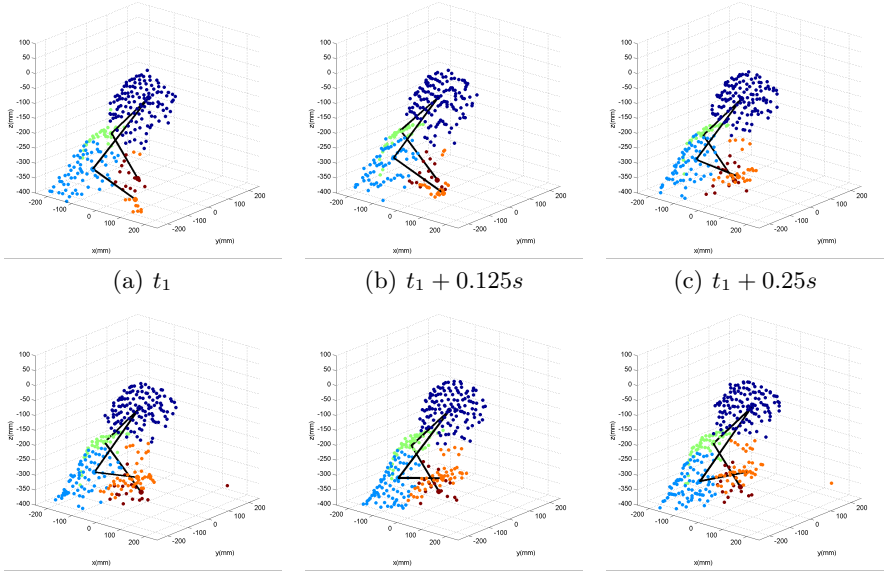
Following the extraction of the bounding box and in order to approximate points relevant to the skeletal structure of the upper body, a modified version of k-means is used to partition the point-cloud into five clusters; these clusters are expected to approximate the position of the head, shoulders and hands. It should be underlined that the produced cluster centroids do not necessarily coincide with a specific anatomical part of the human body, and that they depend on the subject's posture, the position of his/her hands and the existence of objects inside the bounding box. Nevertheless, since we are interested in obtaining information regarding the user motion and posture in uncontrolled, real-life environments, the proposed tracking algorithm can eventually approximate to some extent the user's head and hands. Therefore, from here on the five clusters of the point cloud will be labelled as head, right-shoulder, left-shoulder, right-hand and left-hand just as a naming convention for explanatory purposes.

Two are the main changes applied herein to the original k-means. First, our clustering method is initialized at each frame using the cluster centroids as produced in the previous frame, instead of using random points. This way, the algorithm converges much faster, since our frame rate is high enough (30 frames per second) to allow only small gesture changes between frames. On the other hand, in order to minimize the effects of the camera's position relatively to the occupant and increase the possibility that the clustering will comply with the expected structure of the human torso, our optimization scheme includes also the position of five fixed points as shown in Fig. 2(b). These points were defined based on the geometry of the human body in the seated position and correspond to the expected positions of head, shoulders and hands. Equation 3 summarizes our clustering optimization scheme for the partitioning of the n points in the

cloud $\mathbf{X}^k = (x_1^k, x_2^k, ...x_n^k)$ of frame k, into five clusters $\mathbf{S}^k = (\mathbf{S}_1^k, \mathbf{S}_2^k, ...\mathbf{S}_5^k)$ based on their centroids $\mu^k = mean(x^k \in S_i^k)$ and the fixed points $\mathbf{Y} = (y_1, y_2, .., y_5)$.

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{5} \sum_{x_j^k \in S_i^k} ||x_j^k - \mu_i^k||^2 * ||x_j^k - y_i||^2 \tag{3}$$

Indicative results of our method are shown in Fig. 3 (calculated point cloud clusters and their centroids); here, it is shown how the calculated centroids follow the movement of the right hand as it approaches the user's head.



(a) t_1 (b) $t_1 + 0.125s$ (c) $t_1 + 0.25s$

(d) $t_1 + 0.375$ (e) $t_1 + 0.5s$ (f) $t_1 + 0.625s$

Fig. 3. Indicative clustering output during eating activity

Finally, the calculated centroid coordinates are used for the definition of the three first features proposed in this study, as shown in Fig. 2(c). The first feature is defined as the average of the distances of the head centroid from the two hand centroids: $f_1 = (d_1 + d_2)/2$, whereas the second regards the distance of the two hands: $f_2 = d_3$. The third feature is defined as the average of the angles between the head, shoulder and hand centroids for both sides: $f_3 = (\theta_1 + \theta_2)/2$. Finally, and in order to represent the relative movement of the user's upper body and the dynamic changes in her/his gesture, three additional features are defined as the standard deviation of 240 frame windows for each one of the above features. In order to validate the potential of these features on discriminating among eating/non-eating activities, ten periods of 10s were randomly selected from the Dataset C described at Section 4, half corresponding to eating activities and half

to seated, non-eating activities (e.g. Reading). From each period, the proposed features were extracted and the Kruskal-Wallis test was employed to test the differences of the feature values distributions between the two classes (i.e. eating vs. non-eating); statistically significant ($p < 0.001$) difference was found for all features.

3 Activity Detection Framework

The activity detector of the present work is based on the theory of Conditional Random Fields; in particular, an HCRF is employed as our activity classifier. Given a set of observations X, Conditional Random Fields compactly represent the conditional probability of a particular label sequence Y through an undirected graphical model, as [24]: $P(Y|X) = (1/Z) \prod_{t=1}^{T} exp(w \times f(t, y_{t-1}, y_t, X))$, where $Z = \sum_Y \prod_{t=1}^{T} exp(w \times f(t, y_{t-1}, y_t, X))$ is a normalization constant and w is the set of weights, representing the parameters that are fitted during training. The weights are multiplied by a vector of computed features $f(t, y_{t-1}, y_t, X)$, which derive from the observations data. The weights-features set represent the potentials $\psi(t, y_{t-1}, y_t, X) = exp(w \times f(t, y_{t-1}, y_t, X))$ of the CRF cliques, which consist of an edge between y_{t-1} and y_t as well as the edges from these two labels to the set of observations X [24].

At this point the basic difference between HMMs and CRFs becomes evident; HMMs model the joint probability of both the labels and observations under the independence assumption of observations, i.e. $P(X, Y)$, whereas CRFs model directly the conditional probability $P(Y|X)$, so as to discriminate between different labels on the basis of not conditionally-independent observations. In practice, CRFs can assign a label to the features derived from observations at each time step. In our case, given for e.g. an one-minute long observation sequence, with observations (i.e. user location) taken at a rate of 1 HZ, we need to find a single label for the whole sequence that best describes the respective activity (i.e. cooking, eating, etc.). Through CRFs, this problem can be approached by finding the Viterbi path under the model and employing majority voting over the labels sequence to assign the dominant label [26]. In [26] however, HCRFs provided an alternative of significant potential to improve effectiveness in discriminating among different gestures based on user hand trajectories. As a basic aim in this work is to discriminate user trajectories on the house floor plan among different daily activities, given the similarities of our problem to the one of discriminating gestures from hand trajectories, it is reasonable to expect that HCRFs could improve performance in our context, as is further explained below.

Hidden Conditional Random Fields employ a set of hidden states to capture the certain underlying structure of each class. In particular, an HCRF models the conditional probability of a class label given a set of observations by [26]:

$$P(y|X, \theta) = \sum_s P(y, \mathbf{s}|X, \theta) = \frac{\sum_{\mathbf{s}} exp(\psi(y, \mathbf{s}, X; \theta))}{\sum_{y' \in Y, \mathbf{s} \in S^m} exp(\psi(y', \mathbf{s}, X; \theta)} \quad (4)$$

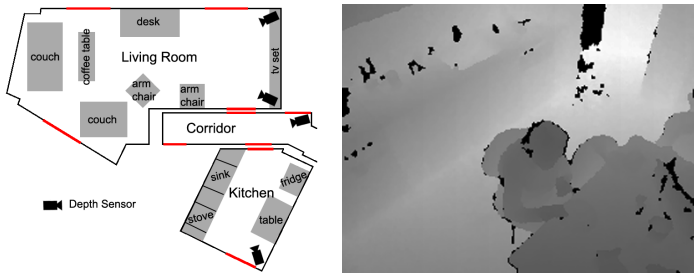
where $\mathbf{s} = s_1, s_2, ..., s_m$, each $s_i \in S$ captures a certain underlying structure of each class and S is the set of hidden states in the model. The potential function $\psi(y', \mathbf{s}, X; \theta) \in \rho$, parametrized by θ (parameter values of the model), measures the compatibility between a label, a set of observations and a configuration of the hidden states. By definition, HCRFs provide a modelling solution that directly addresses the needs of our problem's formulation; given an observations sequence, HCRFs build upon an underlying graphical model that captures temporal dependencies among observations, so as to derive a single label that better corresponds to the input sequence. Considering the simplest case of our specific problem, where the classifier's input is the user's location trajectory during a target activity, the HCRF employs hidden states so as to model dependencies between observations, toward recognizing the target activity being performed.

As such, a HCRF can provide more direct inference for our problem, compared to CRFs, but also to HMMs. For segmented observation sequences, each corresponding to a given activity from a set of M target activities, a set of M HMMs should be employed, each modelling trajectories of a given activity through $P(X, Y) = \prod_{t=1}^{T} P(y_t|y_{t-1}) \times P(x_t|y_t))$; the first term corresponds to pairs of labels and the second pairs each observation to its parent label. During inference, by employing the forward-backward algorithm, the probability that the respective model can produce the input trajectory can be calculated [6]. On the other hand, a single HCRF can be used so as to directly identify the most appropriate label given the input sequence, on the basis of hidden states that formulate cliques on an undirected graph between the observations and labels. Herein, a given hidden state can encode similar characteristics between two different activities that appear in segments of their observation sequences, whereas the ensemble of hidden states corresponding to each full sequence will eventually produce the required differentiation among the different labels. Moreover, long-range dependencies between observations can as well be incorporated in the HCRF, by modifying for e.g. the potential function ψ in Eq. 4, so as to include a window parameter ω that defines the amount of past and future history to be used when predicting the hidden state at each time step [26]; in this study, we follow [26], thus for window size ω, observations from $t - \omega$ to $t + \omega$ are used to compute the input features of the HCRF.

4 Experimental Evaluation

In order to experimentally evaluate our HCRF-based activity recognition framework in real-world activity monitoring scenarios, we first used the two datasets of [17]. These datasets allowed a direct comparison of our HCRF-based framework with the HMM-based approach that was followed in [17] to take place.

Moreover, in order to evaluate our approach of HCRF-based fusion of user location trajectories with information regarding the user's posture and gestures in realistic house settings, we conducted a new data collection experiment, set in a real apartment. More information regarding this dataset will be provided in what follows.



(a) Monitoring area (b) Kitchen depth image

Fig. 4. Apartment experiment setup (Dataset C)

4.1 Datasets Description

The 1^{st} of the datasets from [17](Dataset A) contained user location trajectories data taken from a controlled kitchen environment, where two different persons performed three target activities in the same sequence (cooking, eating, dishwashing) for a total of nine different sessions, while being monitored from a single Kinect camera located at a kitchen's roof-top corner. In this case, four sessions were used for system training and the remaining for evaluation.

The 2^{nd} dataset of [17] (Dataset B) concerned 24/7 monitoring of the resident of a real apartment, through three Kinect sensors, covering the apartment's living-room, the corridor and the kitchen; the resident was monitored for three days, while freely performing the target activities, i.e. cooking, eating, dishwashing and watching TV. In each day, one instance of each target activity occurred. As in the case of Dataset A, user location trajectories derived were available. In this dataset, data from the first two days were used to train the activity detection framework and the data of the third day were used for evaluation.

As mentioned earlier, a new dataset was recorded (Dataset C) from a real apartment scenario, where the resident was monitored for a period of 12 days; using 4 Kinect sensors monitoring the kitchen, corridor and living-room areas. The floor plan of the specific setting is shown in Fig. 4(a), along with the monitoring sensors' location, whereas a sample depth image taken from the apartment's kitchen is shown in Fig. 4(b). The target activities were the ones of Dataset B, with the addition of a non-eating activity, in order to describe different activities, besides eating, that take place in the area of the kitchen table such as reading or talking to the telephone. During the data collection period, 19 cooking, 10 eating, 7 non-eating, 28 dishwashing and 7 watching TV activity instances occurred. From this dataset, we used 2/3 of the activity instances as the train set and the rest were used for evaluation.

In order to train and test the examined classifiers, following the rationale of [9], we split each activity instance of the dataset into one minute-long non overlapping intervals of observations. Each interval was annotated with the label of the respective activity. With respect to each dataset, we obtained the amount of intervals per activity that is shown in Table 1.

Table 1. Number of activity intervals for training and evaluation in each dataset

Activity	Dataset A		Dataset B		Dataset C	
	Training	Evaluation	Training	Evaluation	Training	Evaluation
Cooking	20	25	48	24	37	22
Dish Washing	12	15	16	6	27	12
Eating	12	15	18	8	30	26
Non-Eating	-	-	-	-	22	20
Watching TV	-	-	110	51	40	36
Total	44	55	192	89	156	116

4.2 Results

Our framework's evaluation consisted of three different steps, involving the above datasets from the different settings. In the first step, the effectiveness of the HCRF activity detector was compared to the HMM-based one of [17], in recognizing the target activities by using solely user location trajectories data. Therefore, observations consisted in this step solely of the (x, y) timeseries of user location on the house floor plan. Evaluation at this step was performed on the basis of datasets A and B [17], while the HCRF was trained through a L-BFGS optimizer [14], using 10 hidden states and a 15sec. window.

Table 2. Detection results of the HCRF-based and HMM-based methods on datasets A and B

Activity	Dataset A				Dataset B			
	HCRF		HMM		HCRF		HMM	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Cooking	93,48	82,70	99,34	76,15	92,00	95,83	66,07	92,50
Dish Washing	81,30	100,00	88,71	89,28	100,00	72,72	100,00	45,00
Eating	87,10	87,09	90,32	91,60	85,71	100,00	100,00	90,00
Watching TV	-	-	-	-	98,07	100,00	92,73	100,00
Average	87,30	89,93	92,79	85,67	93,95	92,14	89,70	81,88

As shown in Table 2, the HCRF was found to clearly outperform the HMM-based approach on the B dataset, while both methods produced similar results on the controlled A dataset . This was an expected result, in line with [26], given the capability of HCRFs to better learn common structures among trajectory classes and uncover the distinctive configuration that sets one trajectory class uniquely against others. In fact, our results are also in line with [9], where per timeslice ADL recognition performance was found to increase when discriminative models (CRF) were compared to generative (HMM) ones, in a state-change sensors-based monitoring context. The fact that the HMM-based approach used herein for comparison was already found in [17] to provide ADL recognition

effectiveness at a better or similar level of past related works, such as [9][7] and [6], further underlines the potential of HCRFs to lead into increased ADL recognition effectiveness on the basis of user location trajectories.

Next, the HCRF-based approach was used on the C Dataset, which also included non-eating activities around the kitchen table. As can be seen in Table 3, our method performed fairly well, producing overall precision of 81.9%. However, it is clear that using only the (x, y) timeseries is not sufficient to succesfully discriminate between cooking, eating and non-eating activities that take place in the area of the kitchen table; many non-eating activity instances were detected as either cooking or eating activities.

Table 3. Detection results of the HCRF-based methods on dataset C, using only the (x, y) timeseries of user location on the house floor plan

Activity	Intervals	Detected As					Prec.	Rec.
		Cook	Dish	Eat	Non-eat	TV		
Cooking	22	14	1	1	6	0	63,64	93,33
Dish Washing	12	1	9	1	1	0	75,00	81,82
Eating	26	0	0	18	8	0	69,23	85,71
Non-Eating	20	0	1	1	18	0	90,00	54,55
Watching TV	36	0	0	0	0	36	100,00	100,00
Overall	116	15	11	21	33	36	81,90	81,90

Table 4. Detection results of the HCRF on dataset C, using the (x, y) user location trajectories, user posture and point cloud features of head and hand gestures (f_3), (f_5)

Activity	Intervals	Detected As					Prec.	Rec.
		Cook	Dish	Eat	Non-eat	TV		
Cooking	22	19	2	1	0	0	86,36	90,48
Dish Washing	12	2	10	0	0	0	83,33	71,43
Eating	26	0	0	25	1	0	96,15	86,21
Non-Eating	20	0	2	3	15	0	75,00	93,75
Watching TV	36	0	0	0	0	36	100,00	100,00
Overall	116	21	14	29	16	36	90,52	90,52

Finally, at the third step of evaluation, our proposed methods for estimating (a) user posture and (b) point-cloud features of the user's silhouette expressing head and hand gestures were also involved, so as to provide further features in the HCRF classifier. By trying different feature combinations, the use of the average of the angles between the head, shoulders and hands centroids (f_3), the deviation of the distance between the hands (f_5) and the user posture, in addition to the (x, y) timeseries, produced the best results. As can be seen in

Table 4, the additional features increased the HCRF's performance by 8.6%, leading to precision at 90.52%. It is clear that the additional features improved the detection precision for eating, cooking and non-eating activities around the kitchen table. Specifically, the user posture contributed to the improvement of the discrimination rate between cooking and eating or non-eating activities, as food preparation typically involves standing postures, while the point cloud features of head and hand gestures helped in discriminating between eating and non-eating activities.

A comparison of the three detection frameworks is presented in Table 5, which shows that our proposed method produced significant increase in activity recognition performance, over both the trajectory-only HMM and HCRF.

Table 5. Comparison of activity detection results on dataset C

Activity	HMM		HCRF (x,y)		HCRF (x,y,f_3,f_5,pstr)	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Cooking	85,31	83,44	63,64	93,33	86,36	90,48
Dish Washing	50,04	88,59	75,00	81,82	83,33	71,43
Eating	81,96	58,24	69,23	85,71	96,15	86,21
Non-Eating	62,78	64,10	90,00	54,55	75,00	93,75
Watching TV	89,92	86,28	100,00	100,00	100,00	100,00
Overall	80,07	76,68	81,90	81,90	90,52	90,52

5 Conclusions

This paper introduced a novel framework for automatic detection of domestic ADLs, such as cooking, dishwashing, eating and watching TV, based on the user's 3D point-cloud extracted through depth video recordings. In this context, the contribution of the present study was two-fold. First, HCRFs were introduced in the context of user location trajectories -based ADL detection and were experimentally compared to HMMs. Taking a further step forward, the present work introduced a novel approach to detecting standing-sitting user postures and more importantly, novel features extracted from the user's point-cloud, related to user head and hand gestures. Through experimental evaluation, it was found that HCRFs improved user location trajectories -based ADL recognition effectiveness compared to HMMs, whereas the inclusion of our proposed (a) standing/sitting posture detection method and (b) point cloud features of head and hand gestures led to further significant (at the level of 8%) increase in performance.

Acknowledgements. This work was supported by the Greek, nationally funded, research project "En-NOISIS".

References

1. Aggarwal, J., Xia, L.: Human activity recognition from 3d data: A review. Pattern Recognition Letters (2014)
2. Ahn, I.S., Kim, J.H., Kim, S., Chung, J.W., Kim, H., Kang, H.S., Kim, D.K.: Impairment of instrumental activities of daily living in patients with mild cognitive impairment. Psychiatry Investig 6(3), 180–184 (Sep 2009)
3. Chen, L., Hoey, J., Nugent, C., Cook, D., Yu, Z.: Sensor-based activity recognition. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42(6), 790–808 (2012)
4. Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q.: Human daily action analysis with multi-view and color-depth data. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) Computer Vision ECCV 2012. Workshops and Demonstrations, Lecture Notes in Computer Science, vol. 7584, pp. 52–61. Springer Berlin Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-33868-7_6
5. Daldoss, M., Piotto, N., Conci, N., De Natale, F.G.B.: Learning and matching human activities using regular expressions. In: Image Processing (ICIP), 2010 17th IEEE International Conference on. pp. 4681–4684 (Sept 2010)
6. Duong, T., Phung, D., Bui, H., Venkatesh, S.: Efficient duration and hierarchical modeling for human activity recognition. Artificial Intelligence 173(78), 830–856 (2009)
7. Fleury, A., Vacher, M., Noury, N.: Svm-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results. Information Technology in Biomedicine, IEEE Transactions on 14(2), 274–283 (2010)
8. Hossain, M., Ahmed, D.: Virtual caregiver: An ambient-aware elderly monitoring system. Information Technology in Biomedicine, IEEE Transactions on 16(6), 1024–1031 (2012)
9. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. pp. 1–9. UbiComp '08, 10th Int. Conf., ACM, New York, NY, USA (2008)
10. van Kasteren, T., Englebienne, G., Krse, B.: Hierarchical activity recognition using automatically clustered actions. In: Keyson, D., Maher, M., Streitz, N., Cheok, A., Augusto, J., Wichert, R., Englebienne, G., Aghajan, H., Krse, B. (eds.) Ambient Intelligence, Lecture Notes in Computer Science, vol. 7040, pp. 82–91. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-25167-2_9
11. van Kasteren, T., Englebienne, G., Kröse, B.J.: Activity recognition using semi-markov models on real world smart home datasets. Journal of ambient intelligence and smart environments 2(3), 311–325 (2010)
12. Lawton, M.P., Brody, E.M.: Assessment of older people: Self-maintaining and instrumental activities of daily living. The Gerontologist 9(3 Part 1), 179–186 (1969)
13. Le, X.H.B., Di Mascolo, M., Gouin, A., Noury, N.: Health smart home for elders - a tool for automatic recognition of activities of daily living. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE. pp. 3316–3319 (Aug 2008)
14. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. In: Mathematical Programming. pp. 503–528 (1989)
15. N., M., Y., Y., N., I., J., G., M., O., Yu, W.: Human body contour data based activity recognition. In: Engineering in Medicine and Biology Society (EMBC). pp. 5634–5637. 35th Annual International Conference of the IEEE (July 2013)

16. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 716–723 (June 2013)
17. Papamakarios, G., Giakoumis, D., Votis, K., Segouli, S., Tzovaras, D., Karagiannidis, C.: Synthetic ground truth data generation for automatic trajectory-based adl detection. In: IEEE International Conference on Biomedical and Health Informatics '14 (BHI 2014) (Jun 2014)
18. Park, S., Kautz, H.: Hierarchical recognition of activities of daily living using multiscale, multi-perspective vision and rfid. In: Intelligent Environments, 2008 IET 4th International Conference on. pp. 1–4 (July 2008)
19. Ross, D.A., Osindero, S., Zemel, R.S.: Combining discriminative features to infer complex trajectories. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 761–768. ICML '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1143844.1143940
20. Samet, H.; Tamminen, M.: Efficient component labeling of images of arbitrary dimension represented by linear bintrees. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(4), 579–586 (July 1988)
21. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efficient human pose estimation from single depth images. In: Criminisi, A., Shotton, J. (eds.) Decision Forests for Computer Vision and Medical Image Analysis, pp. 175–192. Advances in Computer Vision and Pattern Recognition, Springer London (2013)
22. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. Computer Vision and Image Understanding 104(23), 210 – 220 (2006), http://www.sciencedirect.com/science/article/pii/S1077314206001093, special Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour
23. Tapia, E., Intille, S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Pervasive Computing, LNCS, vol. 3001, pp. 158–175. Springer (2004)
24. Vail, D.L., Veloso, M.M., Lafferty, J.D.: Conditional random fields for activity recognition. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 235:1–235:8. AAMAS '07, ACM, New York, NY, USA (2007)
25. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 1290–1297 (June 2012)
26. Wang, S.B., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 2, pp. 1521–1527 (2006)
27. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(7), 1264–1277 (July 2009)
28. Zhang, C., Tian, Y.: Rgb-d camera-based daily living activity recognition. Journal of Computer Vision and Image Processing 2(4), 12 (2012)
29. Zhang, H., Parker, L.: 4-dimensional local spatio-temporal features for human activity recognition. In: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on. pp. 2044–2049 (Sept 2011)

30. Zhao, Y., Liu, Z., Yang, L., Cheng, H.: Combing rgb and depth map features for human activity recognition. In: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC): December. pp. 3–6 (2012)
31. Zhou, Z., Chen, X., Chung, Y.C., He, Z., Han, T., Keller, J.: Activity analysis, summarization, and visualization for indoor human activity monitoring. Circuits and Systems for Video Technology, IEEE Transactions on 18(11), 1489–1498 (2008)