Unsupervised discovery of human activities from long-time videos

Salma Elloumi , Serhan Cosar , Guido Pusiol , Francois Bremond , and Monique Thonnat

STARS Team, INRIA Sophia Antipolis - Mediterranee, 2004 Route des Lucioles, BP 93, Sophia Antipolis, France *

Abstract

In this paper, we propose a complete framework based on a Hierarchical Activity Models (HAMs) to understand and recognise Activities of Daily Living (ADL) in unstructured scenes. At each particular time of a long-time video, the framework extracts a set of space-time trajectory features describing the global position of an observed person

*E-mail: {salma.zouaoui-elloumi,serhan.cosar}@inria.fr,pusiol@stanford.edu, {francois.bremond,monique.thonnat}@inria.fr

and the motion of his/her body parts. Human motion information is gathered in a new feature that we call Perceptual Feature Chunks (PFC). The set of PFC is used to learn, in an unsupervised way, particular regions of the scene (topology) where the important activities occur. Using topologies and PFCs, we break the video into a set of small events (*Primitive Events*) that have a semantic meaning. The sequences of *Primitive Events* and topologies are used to construct hierarchical models for activities. The proposed approach has been experimented in the medical field application to monitor patients suffering from Alzheimer and dementia. We have compared our approach with our previous study and a rule-based approach. Experimental results show that the framework achieves better performance than existing works and has a potential to be used as a monitoring tool in medical field applications.

1 Introduction

Nowadays, there are many applications (such as surveillance, human-computer interaction, etc.) that require an efficient and accurate analysis of human activities using video input. For example, in the medical field, the behaviour of patients (e.g. suffering from dementia or Alzheimer disease) needs to be studied on a long-period of time (days and weeks) in order to help medical staff (doctors, carers and nurses) to understand the difficulties of patients and propose solutions that can ameliorate their daily living conditions [3].

Modelling and recognising activities is a rising field in computer vision and machine learning. Recent approaches [10, 26] address the problem of detecting complex daily activities using egocentric wearable cameras which enable to have a close view and see the objects in their natural positions. However, a wearable camera can be very intrusive for the user, especially for people suffering from dementia. Visual information can also be obtained with fixed cameras. The majority of work in activity recognition using fixed cameras addresses short-term actions (i.e. few seconds) in acted footages of posture-defined classes such as “punching” [29, 13]. In order to recognise human activities, scenes need to be analysed from a sequence of frames (low-level task of computer vision) and interpreted (high-level task). The inability of connecting these two levels (high-level and low-level tasks) is called semantic gap problem [31] and its reduction is still a challenging task.

In this paper, we propose a new approach to reduce this gap by constructing, in an unsupervised manner, an intermediate layer between low-level information (tracked objects from video) and high-level interpretation of activity (e.g cooking, eating, sitting). Our method is a novel approach allowing the detection of complex activities with long-duration in an unstructured scene. We have developed a complete vision-based framework that enables to model, discover and recognise activities online while monitoring a patient. Two main contributions of this work are as follows:

1. An intermediate representation of features (the *Primitive Events*) composed of basic activities which structures the person motion with respect to a spatial topology.

2. A hierarchical activity model, that can categorize complex activities using increasing granularity levels of the spatio-temporal structure of basic activities.

In our previous study [28], by using the same tracking and topology learning procedures, we have proposed an unsupervised method that models activities only based on frequency histograms of two features: i) type of primitive events and ii) the direction of local dynamics. One drawback of this method is that the models are characterized without considering the hierarchical links between primitive events. In order to cope with this drawback, in this paper, we have proposed a new activity model, called hierarchical activity models (HAMs) that take into account the hierarchical structure of primitive events. In addition, we have extended the evaluation by using a dataset that includes non-guided activities of daily living (ADL) and demonstrated that,by using HAM, we achieve better performance than existing works.

We start in Section 2 by presenting the related work and previous approaches in the field of activity recognition. An overview of the proposed activity discovery framework is presented in Section 3 . In Section 4, we describe the low-level video processing and the primitive events. We introduce the process of building the hierarchical activity model in Section 5. Experimental results are discussed in Section 6 and the conclusion is presented in Section 7.

2 Related work

Activity analysis and recognition using video is a fast-growing field based on different methods and techniques. The goal of activity recognition is analysing human activities from an unknown video based on the movements of the person. In general, videos are captured either by a fixed camera [13, 12, 35] or by a wearable camera [25, 32, 11, 10, 26]. A complete overview of the previous methods on human activity recognition is proposed in [2] in which the authors emphasize the importance of high-level activity understanding for several important applications, namely those related to ADL.

A major group of previous work in activity recognition includes knowledge and logic-based approaches [23, 16]. For example, authors in [35] proposed a monitoring system for analysis and recognition of human activities. It includes detecting, tracking people and recognising some pre-defined activities using posture information. Three sources of knowledge were exploited: the model of activities, the 3D model of the observed scene, and the 3D model of the mobile object present in the observed scene. In [7], a knowledge-based method is proposed for older people monitoring. Events are modelled as a function of human body context (e.g., sitting, standing, walking), that is obtained from images, and the environment context, which is obtained from accelerometers attached to objects of daily living (e.g., TV remote control or doors use). A rule-based reasoning engine is used for processing, analysing both context types and detect events that fit in rules. While logic-based approach is a natural way of incorporating domain knowledge, for every deployment it requires an extensive enumeration by a domain expert. In addition, there are some methods that utilise Markov logic networks (MLN) to model events using first-order logic in a Markov network [18, 8]. In [18], they represent each target activity as weighted and undirected trees, starting from primitive actions at the bottom to activities at the top. In [8], an MLN is constructed to recognise ADL in a smarthome using non-visual and non-wearable sensors. To overcome the noisy and unreliable observations coming from the sensors, they build logical models can be checked by human and linked to domain knowledge.

Recently, in order to understand long-term activities a particular attention has been given to trajectory-based approaches that utilize the object trajectory information over time. In general, these approaches can be classified into supervised and unsupervised methods. Using a labelled training dataset, supervised methods [15, 20] can build very precise activity models. However, they require large manually labelled training datasets. Also, Hidden Markov Models (HMMs) are applied for the recognition of daily activities [9, 14].

[9] introduces the Switching Hidden Semi-Markov Model (S-HSMM), a two-layered extension of the hidden semi-Markov model (HSMM) for modelling low-level and high-level temporal structures of activities. They show that the proposed S-HSMM performs better than the HSMMs and the HMMs in the recognition of frequent and infrequent activities. A recent trajectory-based approach for human activity recognition [14] combines hierarchical Dirichlet process and HMM to address some limitations of HMM, especially in predicting the number of human motion states in videos. But it requires a lot of computation to obtain the number of motion states. The HMM-based approaches tries to recognise activities by modelling the time-series function of events and learning the parameters of the function using supervised learning techniques. However, recognising complex events, such as "preparing meal", using time sequence is very difficult since the sequential pattern is person-dependent. The unsupervised methods include works such as [17] in which authors learn motion patterns in traffic surveillance videos by using a two-layered trajectory clustering in space and time via fuzzy k-means algorithm. This idea has been extended in [24] and a three-layered clustering is performed on trajectories in order to learn the variations in spatial routes, time duration and speed. Then, the spatio-temporal dynamics of

each cluster is encoded by training HMMs using the most representative examples of clusters. Other methods [4, 6] use dynamic programming based approaches to classify activities. These methods are only effective when time ordering constraints hold. The approach in [27] uses HMM to represent trajectory paths by clustering and captures spatio-temporal patterns in trajectory paths. Clustering is based on finding the number of clusters by checking how well eigenvectors of the trajectory correlation matrix span the subspace. This approach allows high-level analysis of activities for detecting abnormalities in traffic videos. However, since ADL are more complex than traffic dynamics, using only trajectories are not sufficient to capture spatio-temporal modalities of ADL and make distinction between activities (e.g. there will be no difference between “standing next to table“ and ”eating at the table“).

In the literature, there are some methods that use hierarchical models for activity recognition [19, 1, 34]. The described system in [34], extracts features from wearable sensor data and use a two-layered Bayesian network to model the relation between sub-activities and activities. The sub-activities and conditional probabilities are learned from data but the activities are manually specified. A method that uses passive sensors in smart home environment and a two-layered HMM to model relation between sub-activities and activities is proposed in [19]. Similarly, the system learns sub-activities from data by clustering. For high-level activities, a HMM is trained using manually labelled data. In [1], using the trajectories extracted from a fixed camera, a human behavioural analysis system is proposed. Using time delay neural networks, first, trajectories are classified into four groups: walking, running,loitering and stopping. Then, a rule-based fuzzy system is used to infer macro and group behaviours. The disadvantages of this system is that training is required for neural networks and the fuzzy system requires specific rules to recognise activities, which is not an easy task for complex activities. On the contrary, in our method, without the need of manually annotated ground truth, we automatically learn the hierarchical relations between activities and sub-activities in an unsupervised way.

The next section gives an overview of the proposed approach in this paper.

3 Overview of the proposed Activity Discovery framework

The complete framework that we proposed in this paper can recognise long-term activities (hours) in an unsupervised manner and can be used in unstructured scenes. In order to build a hierarchical activity model that characterizes a complex activity, it uses contextual information to create automatically an intermediate structure of a basic activity. This is performed by following steps: (i) long-term videos are processed in order to obtain important information (features) about an observed person (i.e. global positions and the motion of his/her body parts), (ii) features are used to learn the multi-resolution levels of the scene regions (topology), (iii) features and scene regions are combined together to build primitive events which represent a primitive state transitions within regions, (iv) based on the primitive events, activities are discovered and the model of an activity is built, (v) the recognition is performed by comparing similarity between models of activity.

4 Low-level video processing and primitive events

4.1 Low-level video processing

Our low-level processing is based on two phases: extracting Perceptual Feature Chunks and learning Topologies.

4.1.1 Perceptual Feature Chunks

We define the Perceptual Feature Chunks (PFCs) as a set of particular information (i.e. global and local dynamics) associated to human motion in the video. This information is obtained after decomposing the video into short sequences of images (i.e. video chunks) based on the significant changes of human motion (e.g. speed).

The position of a person, is estimated using a set of tracklets which is computed for each video chunk by tracking particular corner points. First, 500 corner points [30] are randomly initialized and tracked over time using KLT [5]. Second, we compute 4 clusters (k-means) of the points with respect to their speed and position, representing static, slow, medium and fast motion.

Finally, we compute the global position p_t of the person at time t, by averaging the centroids of the 3 point clusters (i.e. slow, medium and fast motion).

Due to noise in images, p_t can be unreliable. Therefore, we obtain a smoothed global position $\tilde{p}_f$ by applying a Kalman filter K_1 to p_t in combination with the last n_s smoothed positions:

$$\tilde{p}_t = \frac{1}{n_s + 1}(p_t + \sum_{i=0}^{n_s} K_1(\tilde{p}_{t-i})) \tag{1}$$

The sequence of $\{\tilde{p}_t\}$ represents the global trajectory which is represented in Figure 1-(a) by green points.

We compute the speed of the person s_t at time t as the difference of the position of the person at time t and $t-1$. Similarly, we compute a smoothed speed, $\tilde{s}_t$, by applying a Kalman filter K_2 to s_t, in combination with the last n_s smoothed speeds:

$$\tilde{s}_t = \frac{1}{n_s + 1}(s_t + \sum_{i=0}^{n_s} K_2(\tilde{s}_{t-i})) \tag{2}$$

Finally, the video is decomposed into video chunks by comparing $\tilde{s}_t$ with a threshold.

Consequently, each video chunk is associated with a PFC that includes following attributes : $Departure_{PFC}$, $Arrival_{PFC}$ which are two Gaussian distributions characterizing the position of the person at the beginning and the end of the video chunk. The mean and standard deviation (μ, σ) of the position distributions are computed using the first (or last) n_g points of the global trajectory. $StartFrame_{PFC}$, $EndFrame_{PFC}$ represent the first and last frame number of the video chunk, respectively. $PixelTracklets_{PFC}$ are the pixel-based tracklets used to calculate the global trajectory of the person. An example of $PixelTracklets_{PFC}$ (pink to purple) of a person moving from the armchair to the kitchen is represented in Figure 1-(a). An illustration of the PFC attributes are presented in Figure 1-(b). The feature chunks enable to collect the necessary information for activity understanding and to avoid expensive computational time, especially for long-term activities. The representation contains minimal but important information about the activity in the scene. For instance, we can store the trajectory information of a 4-hour video in less than 14Kb of memory.

4.1.2 The Topology

When a tracked person performs activities, he/she interacts with many objects that can be represented by fixed regions (e.g. the person interacts with the kitchen to prepare meal). We name each set of scene regions a topology (or contextual information) and learn each topology by clustering trajectory points ($\{\tilde{p}_t\}$).

To learn a topology, we use the PFCs associated to one or several people performing activities in the same scene at various time. From this set of sequences, we extract a set of points, that we call $Points_{Seq}$, using the $Departure_{PFC}$ and $Arrival_{PFC}$ of all videos.

$$Points_{Seq} = \{Departure_{PFC}(\mu)\} \cup \{Arrival_{PFC}(\mu)\} \tag{3}$$

We perform k-means clustering [22] over $Points_{Seq}$. The number of clusters represents the level of granularity of the topology, where lower numbers imply smaller number of regions that are wider. Each cluster defines a *Scene Region* (SR). We denote a topology at level l associated with k clusters as $T_l = \left\{SR_0^l, ..., SR_{k-1}^l\right\}$.

We represent a *scene model* as a vector of topologies of different resolution levels: $\{T_l\}$. We build this scene model by calculating 3 levels of topologies that correspond to 5, 10 and 15 clusters. Figure 2 describes the scene model obtained by clustering extracted points in the HOMECARE dataset (described in Section 6), corresponding to high, medium and low-level activities.

4.2 Primitive Events

We propose an intermediate layer called $Primitive\,Events$ that enable to link gradually the extracted features from images (low-level information) to the semantic interpretation of the scene (high-level information).

$Primitive\,Events$ are the events characterizing Perceptual Feature Chunks (section 4.1.1) over a single topology (section 4.1.2). For each person, a sequence of $Primitive\,Events$ is built using the sequence of PFCs and a topology T_l. In practice, we build 3 sequences of $Primitive\,Events$ (for $l = 1$, 2 and 3) for a single video.

$Primitive\,Events$ has 2 attributes, called $Transition_{PE}$ and $LocalDynamics_{PE}$, that contain extracted features and their semantic interpretation.

4.2.1 The $Transition_{PE}$

It describes the movement of a person over the scene by extracting the transition information performed between learned scene regions SR_i^l at one level, l.

The $Transition_{PE}$ is represented as a directed region pair:

$$Transition_{PE} = (StartRegion \rightarrow EndRegion) \tag{4}$$

where *StartRegion* and *EndRegion* are the labels of the nearest SR_i^l (i^{th} scene region from T_l) to the $Departure_{PFC}(\mu)$ and $Arrival_{PFC}(\mu)$ positions.

4.2.2 The $LocalDynamics_{PE}$

The $Transition_{PE}$ can only describe the global motion of the person while he/she performs an activity over the scene (moving from one region to another one or staying in a region). To be able to model finer activities (low-level activities), we compute the $LocalDynamics_{PE}$ attribute that contains finer information (point tracklets) on the movement of the human body parts (hands, arms, torso, etc).

The $LocalDynamics_{PE}$ are obtained by clustering the $PixelTracklets_{PFC}$ (section 4.1.1). For clustering, we use the mean-shift algorithm [33]. In the literature, the methods for tuning the bandwidth of the mean-shift algorithm are not appropriate to compute a finer description of the local motion. Thus, we adapt the mean-shift bandwidth automatically as a function of the global position of the person:

$$h = ||Departure_{PFC}(\mu) - Arrival_{PFC}(\mu)|| \tag{5}$$

where h is the bandwidth window. Figure 3 illustrates five examples of the computed $LocalDynamics$ (green) from the clustering of the $PixelTracklets_{PFC}$ (pink) associated to the following movements: arms up, arms down, join hands, bend down and stretch up. It can be seen in the figure how local dynamics (green tracklets) can capture five activities while the person remains at the same location.

5 Building the Hierarchical Activity Model

5.1 The process of Activity Discovery

The sequences of $Primitive\,Events$ are very informative about the activity occurring in the video. However, a $Primitive\,Event$ can only describe a snapshot of the person motion. In order to provide more meaning, a better representation of the discovered activity is needed.

If a person stays in a region for a certain amount of time, we need to fuse the sequences of $Primitive\,Events$ to obtain one global activity corresponding to all the time he/she stayed in the region. Another kind of activity occurs when the person moves from one region to another. Therefore, we consider two patterns, $Change$ and $Stay$, to describe the two types of activity:

- The $Stay$ pattern characterizes an activity occurring within a single topology region like "at.region.P", and it is defined as a maximal subsequence of $Primitive\,Events$ with the same $Transition_{PE}$:

$$Stay_{P-P} = (P \rightarrow P)^{+} \tag{6}$$

- The $Change$ pattern describes the transition of the person between regions like "changing.from.P.to.Q" which is composed of a single $Primitive\,Event$:

$$Change_{P-Q} = (P \rightarrow Q), P \neq Q \tag{7}$$

We define a discovered activity (DA) at a level l as an extracted $Stay_{P-P}$ or $Change_{P-Q}$ pattern:

$$DA^{l}_{P-Q} = Stay_{P-P}|Change_{P-Q} \tag{8}$$

The process of activity discovery is performed over the three granularity levels ($l = 1, 2, 3$) by using the three sequences of $Primitive\,Events$. Therefore, based on the hierarchy of the scene regions, the discovered activities are also classified to coarse, medium and fine and each of them is a sub-activity of an activity at a coarser resolution.

In the following sections, we replace $P - Q$ and $P - P$ by the index s that represents the semantic of an activity. Each activity are mapped to a colour on the graphical interface to categorize the activities in the video. Figure 4 shows the coloured segments representing the discovered activities at three levels of resolution. Same colours correspond to the same activity at each resolution level.

5.2 The Hierarchical Activity Model

5.2.1 Definition of the model

We represent the model of an activity as a tree of nodes that is obtained by merging the set of $\{DA_s^{l=1,2,3}\}$ (s is the semantics of the activity) and has a hierarchical structure based on the three levels of granularity (i.e. $\{N^{l=1}, \{N_i^{l=2}\}_{1\leq i\leq n}, \{N_j^{l=3}\}_{1\leq j\leq m}\}$). The tree of nodes represents how different activities and sub-activities are connected to each other thanks to a set of $attributes$ and $sub-attributes$ obtained from the properties such as type, duration, etc. In other words, a node N is characterized by $attributes$ and $sub-attributes$:

- The $attributes$ is a set of parameters over the DAs at the current level l that characterizes the node N^l.
- The $sub-attributes$ constitutes the set of parameters that characterizes the attributes of the sub-nodes N_i^{l+1}, where i is the index of the child node of N^l.

5.2.2 Learning phase of the model

For a selected instances of the same discovered activities DA_s^l (e.g. $s=$ "cooking"), we learn the model of activity by constructing a tree of nodes where each node of level l is built from the set of discovered activities that are at the same resolution level l, $\{DA_{s_1}^l, DA_{s_2}^l, ..., DA_{s_n}^l\}$ where $s_1, s_2, ..., s_n$ are parts of s (i.e. sub-activities of cooking). An example of the constructing process of a tree of nodes from three sequences of discovered activities classified from the coarser to the finer one is illustrated in Figure 5-(a). We construct an independent model for each type of discovered activity. In the following subsections, we describe the parameters of $attributes$ and $sub-attributes$.

The attribute of a node For a node N^l, we define 3 attributes to describe temporal and spatial properties of a node:

- *Type*: it is adopted from the DAs composing a node. For a node N, $type_N = type_{DA_s}$
- *Instances*: the amount of training instance of activities composing a node.
- *Duration*: a Gaussian distribution $N(\mu_d, \sigma_d^2)$ describing the temporal duration of the training instances.
- *Histogram of Local Dynamics* $H(\theta)$: is a histogram that characterizes the length and the angle of local motion. As it is presented in Figure 5-(b), the length is the magnitude of the local motion vector and the angle is orientation of the vector with respect to x-axis, which is discretised into 8 bins.

The sub-attribute of a node The sub-attributes enable us to get information from the child nodes. To compute the sub-attributes of a node, we use the attributes of its child nodes. For a node N^l, we define two sub-attributes named $mixture_{sub-activity}$ and $timelapse_{sub-activity}$ which aim at describing two properties of the child nodes N_i^{l+1} of N^l:

1. $mixture_{sub-activity}$: Describes the amount of time a child node with the same $Type$ appears. It is represented as a mixture of Gaussians (MOG) of ($\theta_{type}^{mixture}$) with the following parameters:

   - K, is the total number of components (Gaussians) and equal to the number of unique $Type$s
   - O, is the total number of discovered activities at level l (DA^l).
   - $w_{q=1...K}$, is the prior probability of the component q. It is equivalent to the weight of each Gaussian in the MOG. It is computed based on the number of appearances of the nodes with the same $Type$:

$$w_q = \frac{\sum_{p=1}^{O} \delta(Type_{N_p^{l+1}}, Type)}{O} \tag{9}$$

   Then, $\theta_{type}^{mixture} = \sum_{q=1}^{K} w_q * N(\mu_q, \sigma_q)$ where μ_q is calculated by the training instances of all child nodes with the same $Type$:

$$\mu_q = \frac{\sum_{p=1}^{O} Instances_{N_p^{l+1}} * \delta(Type_{N_p^{l+1}}, Type)}{\sum_{p=1}^{O} \delta(Type_{N_p^{l+1}}, Type)} \tag{10}$$

2. $timelapse_{sub-activity}$: Represents the temporal distribution of child nodes. For an activity, it describes the expected temporal duration of its

sub-activities. $timelapse_{sub-activity}$ is also represented by a MOG of ($\theta_{type}^{timelapse}$). The parameters of $timelapse_{sub-activity}$ are similar to previous sub-attribute $mixture_{sub-activity}$.

5.2.3 Recognition phase of the model

For a new unseen video dataset, we aim at recognising activities in an unsupervised way. The task is achieved by measuring the similarity between reference activity models that are learned for each type of discovered activity using unlabelled training videos and a test activity model that is obtained from the discovered activities of the new video.

First, a new sequence of Perceptual Feature Chunks are computed for the new video. Second, using three levels of topology learned from training videos, we create new $Primitive\,Events$. Thereby, $Transition_{PE}$ of new $Primitive\,Events$ are matched with the $Transition_{PE}$ of $Primitive\,Events$ used in training. Third, the activity discovery process is performed with the new $Primitive\,Events$ and a new sequence of discovered activities are computed. Fourth, for each type of discovered activity of the new video, an activity model is built as explained in Section 5.2.2. Finally, we compute a score between the new model and learned models and classify the activity by assigning the label of the best match.

To compute a similarity score between two activity models, we define a metric in a recursive manner. At each level of the model, we calculate a similarity score by computing the Euclidean distance between attributes and sub-attributes of the nodes of two models at that level and append the similarity score obtained from the finer level. Since the range of attributes vary, we have normalised the distances. This recursive procedure give us the opportunity to have a similarity score at the root node that measure the similarity of the models at all levels.

6 Experimental results

We have tested the proposed framework on three datasets. Each video in the dataset contains one person and is recorded using a monocular video camera with 640×480 pixels of resolution. The size of the person is about 50×150 pixels. The three datasets are as follows:

a) HOMECARE dataset: It consists of a set of 7 videos associated to seven people performing everyday activities in an apartment (activities are listed in Table 1-(a)). The apartment, which has a size of $42m^2$, is an experimental laboratory set up under the national project Gerhome. Each video is of 4-hour length. An overview of the scene and a sample of activities are presented in Figure 2-(a) and Figure 6, respectively.

b) HOSPITAL dataset: It includes a set of 4 videos associated to 4 patients and recorded in a hospital room, which has a size of $32m^2$, while the patients are visiting their doctors. The patients perform some guided activities from a medical protocol. Figure 8-(e) shows the overview of the hospital room and Figure 7 describes the set of activities that we aim to recognise. Each video lasts 1 hour.

c) CHU dataset: It consists of a set of 30 videos associated to 30 patients in the same room of HOSPITAL dataset. This dataset is more challenging than HOSPITAL dataset, since the person performed a non-guided activities of daily livings. The activities of interest are given in Figure 8-(a-d).

The classification results for HOMECARE and HOSPITAL datasets are based on leave-one-out cross validation. The evaluation is performed by learning the scene (Section 4.1.2) and activity (Section 5.2.2) models from the training videos and by recognising activities in a test video. In the HOMECARE dataset, the scene and activity models are learned after processing 6 videos. The remaining video from the HOMECARE dataset is used for recognition procedure. First, activities are discovered in the remaining video using the set of extracted $Primitive\,Events$ associated to the person and the scene model learned from 6 videos. Then, for each discovered activity, an activity model is created and compared with the activity model learned from training videos (Section 5.2.3). Similarly, in the HOSPITAL dataset the scene and activity models are learned using 3 videos and one video is selected to recognise activities. For the CHU dataset, we have randomly selected 10 videos for learning the scene and activity models. The remaining videos are used to recognise activities.

To evaluate the framework, we have used True Positive (TP), False Positive (FP), False Negative (FN) and calculated $Sensitivity$ and $Precision$ as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

An example of learned scene model for the HOMECARE dataset is represented in Figure 2. The performance of our framework for HOMECARE, HOSPITAL and CHU datasets are displayed in Table 1-(a), Table 1-(b) and Table 1-(c), respectively. The recognition results of the proposed framework are obtained by comparing with manually annotated ground truths. It can be seen from Table 1 that the proposed method proved to be very reliable for activities between two areas (e.g.“armchair” to “table” in Table 1-(a), ”exercise 1” in Table 1-(b) and “office desk” to “drugs desk” in Table 1-(c)). The proposed method is also good at recognising activities occurring in one area (e.g. “reading in the armchair” in Table 1-(a), “preparing coffee” in Table 1-(b) and “preparing drugs” in Table 1-(c)). Thanks to the proposed hierarchical model of activities (section 5.2), we are able to recognise different activities performed within a particular area. For example, the Table 1-(a) shows that the system has detected two different activities (standing and reading) when the person is in the armchair area (2^{nd} area for $k = 5$ in Figure 2-(b)). The discovery and distinction between two different activities occurring in the same area are possible thanks to the use of the local dynamics (section 4.2.2). This can also be seen in Table 1-(b) where the two different exercises (Up/down and Balance) occurring in the same area are not confused.

The reason of failure in detecting an activity (i.e. False Negative) is the failure in motion detection. The process of trajectory extraction described in Section 4.1 sometimes fails to track people. Because of the inadequate trajectory information, we have many FNs in CHU dataset. For HOMECARE and HOSPITAL dataset, the false detection of an activity (i.e False Positive) usually happens when the person stops an activity without changing his/her place (e.g. the person stays still for a while at the end of eating activity). Recognizing non-guided ADL is more challenging. In CHU dataset, we have high FP rates because some of the learned zones are very close to each other. For instance, for the actions of "preparing tea“ and “talking on the phone”, we obtain a high rate of FP, because the zones where the actions occur in are very close to each other. Therefore, these actions are misclassified.

Considering the results in Table 1, it can be seen that the framework achieves a high rate of True Positive and a low rate of False Negative. In total, majority of the performed activities are recognised by the framework.

The concept of primitive events together with hierarchical activity models also enables us to handle the problem of occlusion. In the case of occlusion, as long as some motion is detected on the visible body parts we could be able to create primitive events and, then, activity models. In severe cases, our framework may miss some instances of primitive events. In fact, occlusion is one of the reasons that causes FNs in Table 1. However, since we statistically learn activity models, it is still possible to build the model from the discovered activities (not occluded) and perform recognition.

We have also analysed the effect of the number of clusters in topology learning phase (k parameter in Section 4.1.2). We have tested the performance of the proposed method by selecting different number of clusters. Table 2 shows the average sensitivity and precision values obtained by selecting the number of clusters as 5,10,15; 7,10,15 and 7,11,16 in CHU dataset. It can be seen that the number of clusters does not significantly affect the recognition performance of the framework.

In Table 3, we have compared the proposed framework with our previous work described in [28] and a rule-based method proposed in [35] where the activities are manually modelled by setting rules and constraints. In [35], they cannot differentiate finer activities inside an area (e,g, “sitting at the table” and “eating“) and recognition performance for some activities are not presented. Thus, for the method in [35], we have given only the results they have presented and the accumulated recognition rate for merged activities. The bold values in the table show the best result for each activity class. It can be seen that for all activity classes the proposed method gives a better rate of sensitivity and precision compared to the method in [35]. Unlike in [35], it can be seen that the HAM is capable of differentiating finer activities. Compared to the method in [28], the proposed HAM enables us to enhance the recognition results. In two activities (”eating“ and ”preparing meal“) we achieve better sensitivity and precision rates and in two activities (”inside bathroom“ and ”from armchair to table“) we achieve better precision rates.

7 Conclusion

In this paper, we have proposed a complete unsupervised framework for discovering, modelling and recognising activities of daily living using a fixed camera in an unstructured scene. This framework includes all steps from the low-level processing to the semantic interpretation of the motion in the scene. Global and local human features are extracted from the video and used to learn meaningful areas (topologies) of the scene in an unsupervised way. Combining global and local features with topologies enables us to build primitive events in the video at different levels of resolution. Following these steps, we have proposed a new model for representing activities: Hierarchical Activity Model which benefits from the multi-resolution structure in primitive events.

The contributions of the framework are twofold: primitive events and hierarchical activity models. To bridge the semantic gap we have proposed an intermediate layer of primitive events which are used to link semantics with perceptual information. Thanks to this intermediate layer, the proposed method overcomes the problem of manually describing the target activities. The hierarchical activity model give us the opportunity to categorize complex activities using increasing granularity levels of the spatio-temporal structure of basic activities.

This framework has been successfully tested for recognising ADL by experimenting in an apartment and in a hospital room. Although there are some missed activities because of failure in detecting finer motion, the experimental results show that the framework is a successful system that can automatically discover, learn and recognise ADL. In addition, it can be observed that the framework can be used in medical applications in order to monitor older persons suffering from Alzheimer or dementia. The statistical information in HAM provides an important data to learn the normal behaviour models and life pattern of people. Hence, the change in behaviour models can be easily detected and used to evaluate the status of people. We believe that by using motion descriptors such as HoG and HoF [21] we can capture finer motion in the video and obtain better performance.

The framework can also be used in many other fields such as video surveillance of metros and airports. Our future work is going to be the extension of our framework to detect abnormal activities in such applications. In addition, we are going to test our framework in online-learning mode by updating (or creating) the zones and activity models in time with new trajectory information.

References

[1] G. Acampora, P. Foggia, A. Saggese, and M. Vento. Combining neural networks and fuzzy systems for human behavior understanding. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 88–93, 2012.

[2] JK Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[3] AJ Bharucha, C Atkeson, D Chen, et al. Caremedia: Automated video and sensor analysis for geriatric care. In *Annual Meeting of the American Association for Geriatric Psychiatry*, 2006.

[4] Aaron F. Bobick and Andrew D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, 1997.

[5] Jean-Yves Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 2001.

[6] Simone Calderara, Rita Cucchiara, and Andrea Prati. Detection of abnormal behaviors using a mixture of von mises distributions. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 141–146. IEEE, 2007.

[7] Yuanyuan Cao, Linmi Tao, and Guangyou Xu. An event-driven context model in elderly health monitoring. In *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, UIC-ATC '09.*, pages 120–124, July 2009.

[8] Pedro Chahuara, Anthony Fleury, Franois Portet, and Michel Vacher. Using markov logic network for on-line activity recognition from non-visual home automation sensors. In *Ambient Intelligence*, volume 7683 of *Lecture Notes in Computer Science*, pages 177–192. Springer Berlin Heidelberg, 2012.

[9] T.V. Duong, H.H. Bui, D.Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 838–845 vol. 1, 2005.

[10] Alireza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1226–1233. IEEE, 2012.

[11] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ECCV'12, pages 314–327, Berlin, Heidelberg, 2012. Springer-Verlag.

[12] Anthony Fleury, Michel Vacher, and Norbert Noury. Svm-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):274–283, 2010.

[13] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, and Deva Ramanan. Computational studies of human motion: part 1, tracking and motion synthesis. *Found. Trends. Comput. Graph. Vis.*, 1(2-3):77–254, 2005.

[14] Qingbin Gao and Shiliang Sun. Trajectory-based human activity recognition with hierarchical dirichlet process hidden markov models. In *IEEE*

China Summit International Conference on Signal and Information Processing (ChinaSIP), pages 456–460, July 2013.

[15] Shaogang Gong and Tao Xiang. Recognition of group activities using dynamic probabilistic networks. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 742–749. IEEE, 2003.

[16] Somboon Hongeng, Ram Nevatia, and Francois Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, 2004.

[17] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, Dan Xie, Tieniu Tan, and Steve Maybank. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, 2006.

[18] Gowun Jeong and H.S. Yang. Context-aware activity recognition by markov logic networks of trained weights. In *16th International Conference on Virtual Systems and Multimedia (VSMM), 2010*, pages 5–12, Oct 2010.

[19] Tim L.M. Kasteren, Gwenn Englebienne, and Ben J.A. Krose. Hierarchical activity recognition using automatically clustered actions. In *Ambient Intelligence*, volume 7040 of *Lecture Notes in Computer Science*, pages 82–91. Springer Berlin Heidelberg, 2011.

[20] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Proceedings of Ninth IEEE International Conference on Computer Vision (ICCV)*, pages 432–439. IEEE, 2003.

[21] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[22] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[23] Gérard Medioni, Isaac Cohen, François Brémond, Somboon Hongeng, and Ramakant Nevatia. Event detection and analysis from video

streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889, 2001.

[24] B.T. Morris and M.M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2287–2301, Nov 2011.

[25] Matthai Philipose, Kenneth P Fishkin, Mike Perkowitz, Donald J Patterson, Dieter Fox, Henry Kautz, and Dirk Hahnel. Inferring activities from interactions with objects. *Pervasive Computing, IEEE*, 3(4):50–57, 2004.

[26] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854. IEEE, 2012.

[27] Fatih Porikli. Learning object trajectory patterns by spectral clustering. In *IEEE International Conference on Multimedia and Expo (ICME'04)*, volume 2, pages 1171–1174. IEEE, 2004.

[28] Guido Pusiol, Francois Bremond, and Monique Thonnat. Unsupervised discovery, modeling, and analysis of long term activities. In *Computer*

Vision Systems, volume 6962 of *Lecture Notes in Computer Science*, pages 101–111. Springer Berlin Heidelberg, 2011.

[29] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 32–36. IEEE, 2004.

[30] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'94).*, pages 593–600. IEEE, 1994.

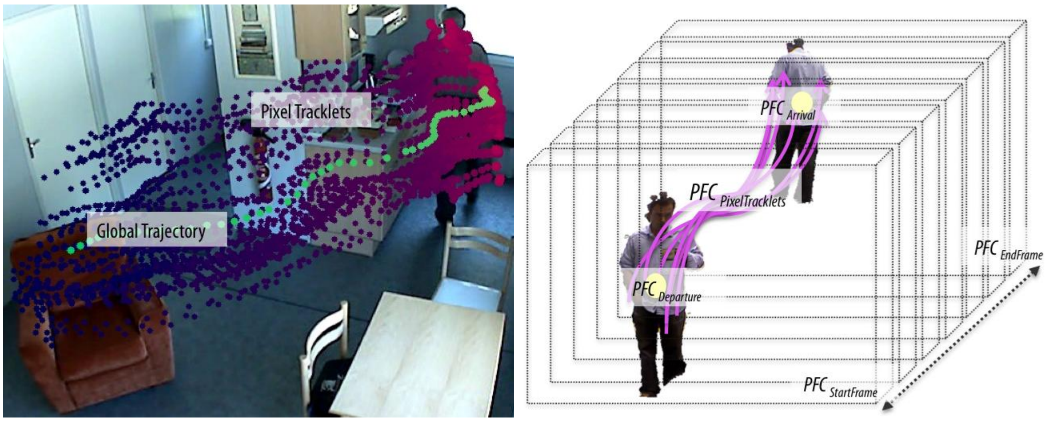
[31] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[32] Ekaterina H. Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 17–24. IEEE, 2009.

[33] Raghav Subbarao and Peter Meer. Nonlinear mean shift over riemannian manifolds. *International Journal of Computer Vision*, 84(1):1–20, 2009.

[34] A. Subramanya, A. Raj, J. Bilmes, and D. Fox. Hierarchical models for activity recognition. In *IEEE 8th Workshop on Multimedia Signal Processing*, pages 233–237, 2006.

[35] Nadia Zouba, Francois Bremond, and Monique Thonnat. An activity monitoring system for real elderly at home: Validation study. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 278–285. IEEE, 2010.



(a) Global trajectory and Pixel tracklets (b) PFC features

Figure 1: Global trajectories (green) and Pixel Tracklets (purple to pink) to construct Perceptual Feature Chunks.

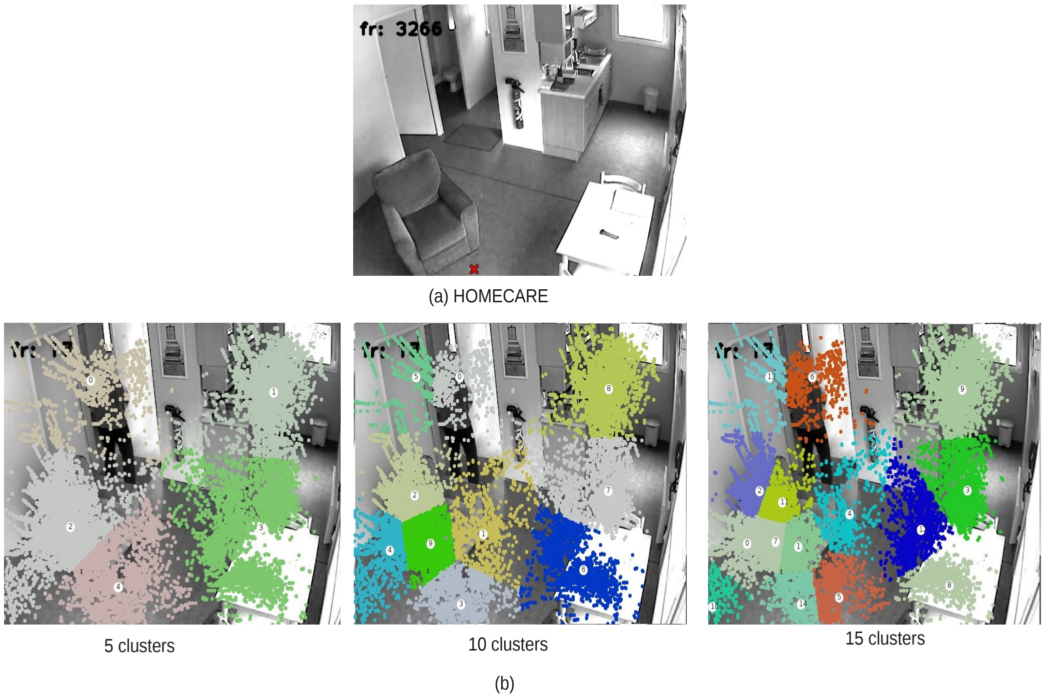


Figure 2: (a) The empty scene for HOMECARE. (b) Example of the scene model with $l = 1$, 2 and 3 obtained by k-means clustering ($k = 5$, 10 and 15) for HOMECARE dataset described in section 6.

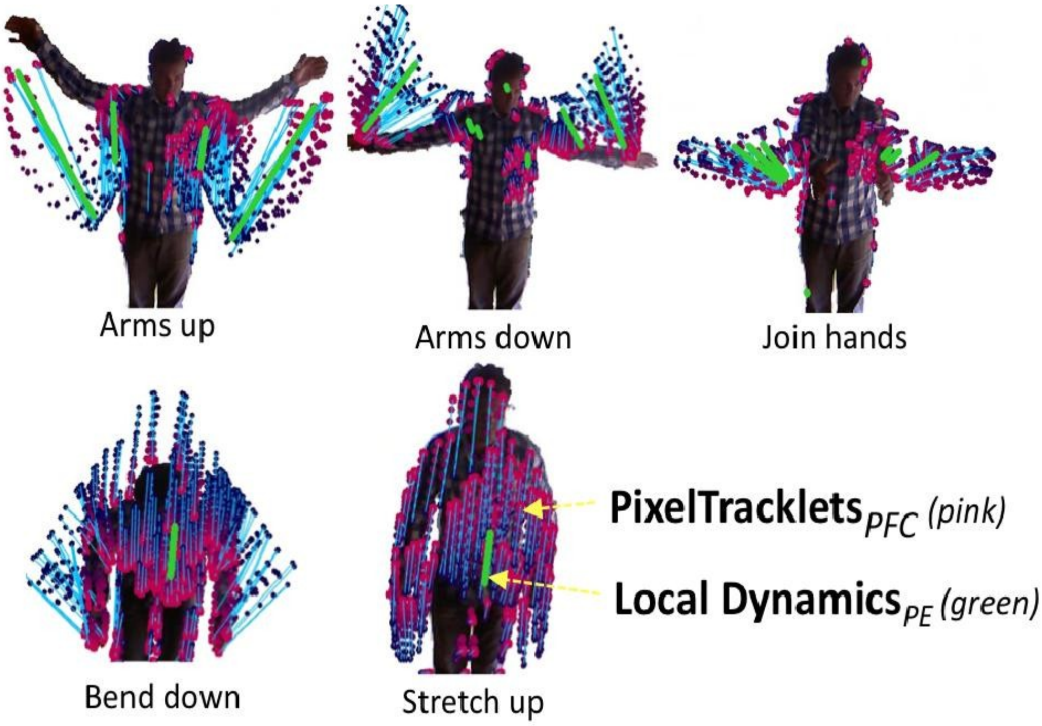


Figure 3: Example of the abstraction of $PixelTracklets_{PFC}$ (pink) into $LocalDynamics_{PE}$ (green). Each, $LocalDynamics_{PE}$ is displayed as a strait line corresponding to the start and end points of an abstracted tracklet (blue line).

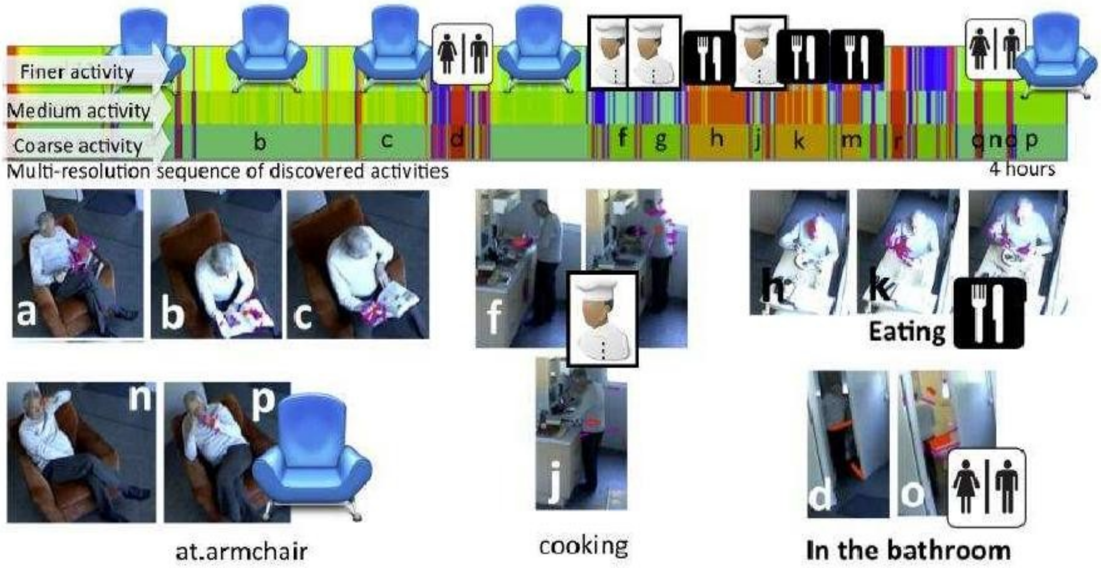
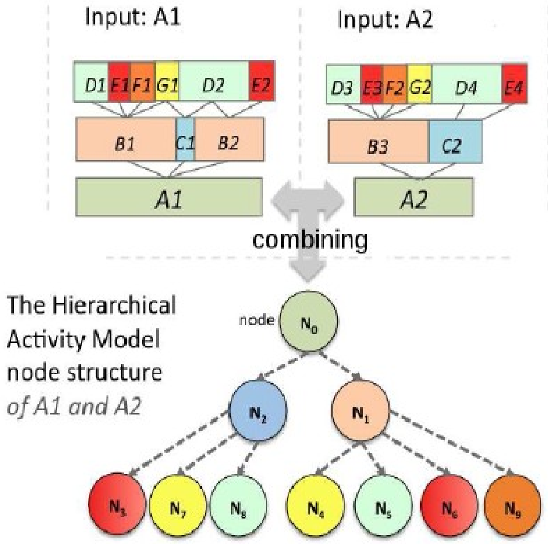
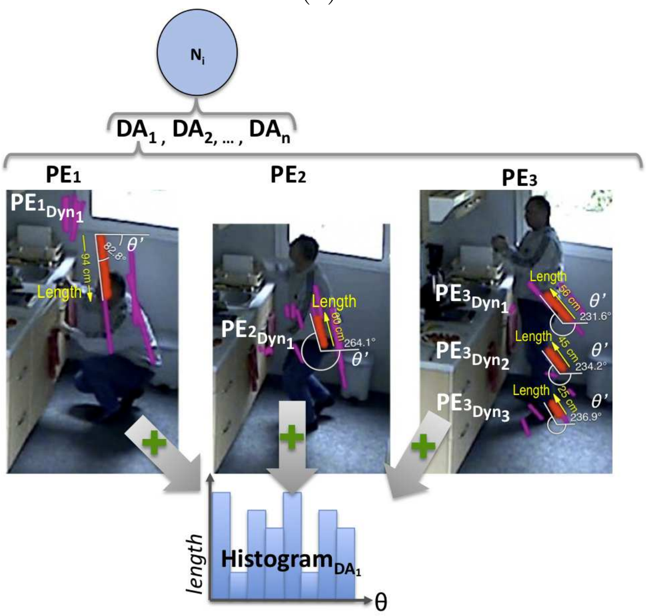


Figure 4: Example of discovered activities (coloured segments) for 4 hours video of one person performing everyday activities. 5 actions in the armchair (a, b, c, n and p), 3 cooking (f, g and j) and eating (h, k and m) actions and 2 actions in the bathroom (d and o) are discovered.



(a)



(b)

Figure 5: (a) Hierarchical Activity Model (HAM) constructed from all DA_s^l (i.e.$\{A_i\}_{1\leq i\leq 2}$ for level 1, $\{B_i\}_{1\leq i\leq 3}$ and $\{C_i\}_{1\leq i\leq 2}$ for level 2 and $\{D_i\}_{1\leq i\leq 4}$, $\{E_i\}_{1\leq i\leq 4}$, $\{F_i\}_{1\leq i\leq 2}$ and $\{G_i\}_{1\leq i\leq 2}$ for level 3 (b) *Histogram of Local Dynamics* attribute computed from the set of discovered activities.

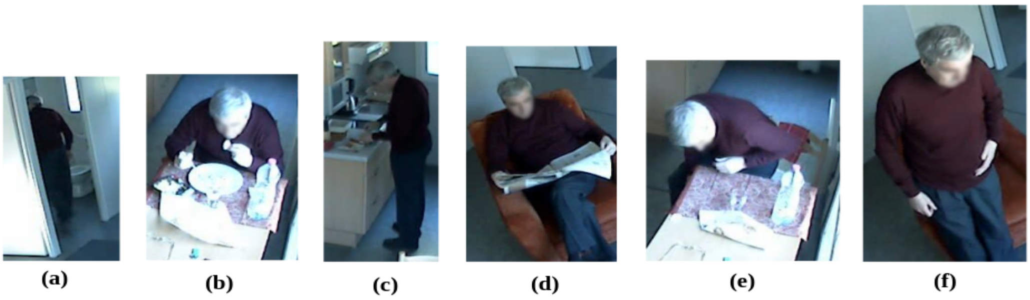


Figure 6: Everyday activities in HOMECARE dataset: (a) In the bathroom, (b) Eating, (c) Preparing meal, (d) Reading in the armchair, (e) Sitting at eating place, (f) Standing at armchair.

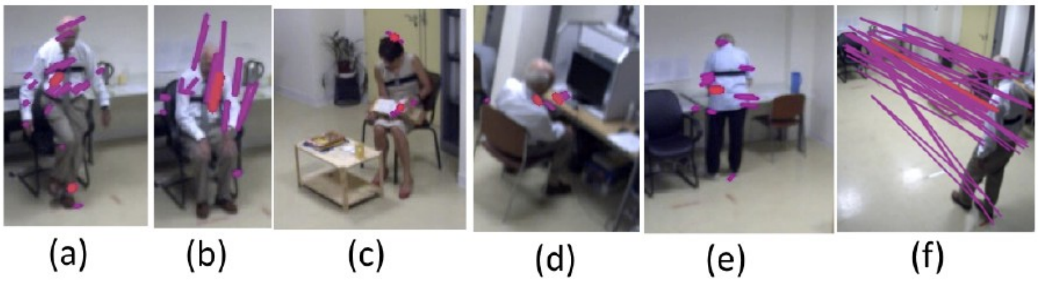


Figure 7: Guided activities in HOSPITAL dataset (a) Balance: Standing on one foot at at time, (b) Up/Down: Standing and sitting down in a continuous way, (c) Reading at the table, (d) At the computer, (e) Preparing coffee, (f) Exercise1/Exercise2: moving from the chair to a marked position and coming back.

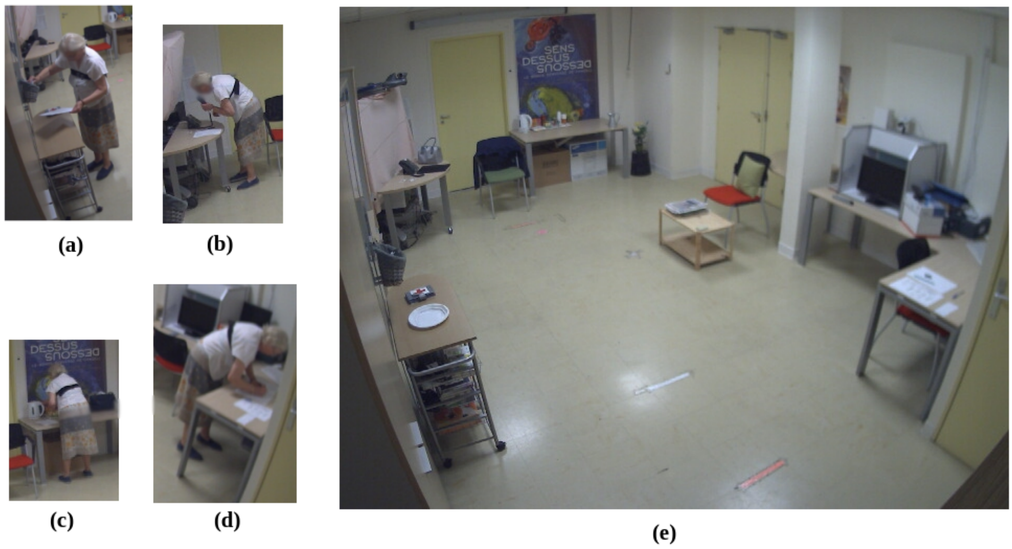


Figure 8: Non-guided activities in CHU dataset: (a) Preparing drugs, (b) Talking on the phone, (c) Preparing tea, (d) Paying bill. (e) Overview of the hospital room in HOSPITAL and CHU datasets.

Table 1: Recognition results for (a) HOMECARE, (b) HOSPITAL and (c) CHU datasets.

	TP	FP	FN	Sensitivity (%)	Precision (%)
Eating	31	1	0	100	96.87
Reading in the armchair	24	5	0	100	82.75
Preparing meal	54	2	1	98.18	96.42
Standing at armchair	11	2	0	100	84.61
Sitting at eating place	8	0	1	88.89	100
Inside the bathroom	14	2	0	100	87.5
From armchair to table	32	2	0	100	94.11
From armchair to kitchen	15	1	0	100	93.75

(a)

	TP	FP	FN	Sensitivity (%)	Precision (%)
Balance	3	0	0	100	100
Up/Down	3	0	0	100	100
Reading at the table	10	1	1	90.91	90.91
Preparing coffee	7	1	0	100	87.5
At the computer	6	1	0	100	85.71
Exercise 1	3	0	0	100	100
Exercise 2	3	0	0	100	100

(b)

	TP	FP	FN	Sensitivity (%)	Precision (%)
Preparing drugs	21	9	1	95.45	70
Talking on the phone	37	12	4	90.24	75.51
Preparing tea	53	11	10	84.12	82.81
Paying bill	40	8	9	81.63	83.33
From office desk to drugs desk	4	1	0	100	80
From drugs desk to tea desk	7	1	1	87.5	87.5

(c)

Table 2: Average sensitivity and average precision for three different level of clusters for CHU dataset.

Clusters	Average sensitivity (%)	Average precision (%)
5,10,15	87.74	77.56
7,10,15	84.06	75.99
7,11,16	80.43	78.37

Table 3: Comparison of recognition rates between the approach in [35], in [28] and the proposed method (specified as "HAM") for HOMECARE dataset.

	[35]		[28]		HAM	
	Sensitivity (%)	Precision (%)	Sensitivity (%)	Precision (%)	Sensitivity (%)	Precision (%)
Eating	78.26	81.81	81	76	**100**	**96**
Sitting at eating place			**88.88**	**100**	**88.88**	**100**
Reading in armchair	85.96	80.32	**100**	**85.71**	**100**	82.75
Preparing meal	80	57.14	88	85	**98.18**	**96.42**
Standing at arm chair	-	-	**100**	**84.61**	**100**	**84.61**
Inside the bathroom	-	-	**100**	77.78	**100**	**87.5**
From armchair to table	-	-	**100**	88.89	**100**	**94.11**
From armchair to kitchen	-	-	**100**	**93.75**	**100**	**93.75**