



**HAL**  
open science

## Unsupervised discovery of human activities from long-time videos

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

► **To cite this version:**

Salma Elloumi, Serhan Cosar, Guido Pusiol, Francois Bremond, Monique Thonnat. Unsupervised discovery of human activities from long-time videos. IET Computer Vision, 2015, pp.1. hal-01123895

**HAL Id: hal-01123895**

**<https://inria.hal.science/hal-01123895v1>**

Submitted on 5 Mar 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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

5 of patients (e.g. suffering from dementia or Alzheimer disease) needs to be  
6 studied on a long-period of time (days and weeks) in order to help medical  
7 staff (doctors, carers and nurses) to understand the difficulties of patients  
8 and propose solutions that can ameliorate their daily living conditions [3].

9

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

23

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

- 32 1. An intermediate representation of features (the *Primitive Events*) com-  
33 posed of basic activities which structures the person motion with re-  
34 spect to a spatial topology.
- 35 2. A hierarchical activity model, that can categorize complex activities  
36 using increasing granularity levels of the spatio-temporal structure of  
37 basic activities.

38 In our previous study [28], by using the same tracking and topology learning  
39 procedures, we have proposed an unsupervised method that models activi-  
40 ties only based on frequency histograms of two features: i) type of primitive  
41 events and ii) the direction of local dynamics. One drawback of this method

42 is that the models are characterized without considering the hierarchical links  
43 between primitive events. In order to cope with this drawback, in this paper,  
44 we have proposed a new activity model, called hierarchical activity models  
45 (HAMs) that take into account the hierarchical structure of primitive events.  
46 In addition, we have extended the evaluation by using a dataset that includes  
47 non-guided activities of daily living (ADL) and demonstrated that, by using  
48 HAM, we achieve better performance than existing works.

49

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

## 56 **2 Related work**

57 Activity analysis and recognition using video is a fast-growing field based  
58 on different methods and techniques. The goal of activity recognition is

59 analysing human activities from an unknown video based on the movements  
60 of the person. In general, videos are captured either by a fixed camera  
61 [13, 12, 35] or by a wearable camera [25, 32, 11, 10, 26]. A complete overview  
62 of the previous methods on human activity recognition is proposed in [2]  
63 in which the authors emphasize the importance of high-level activity under-  
64 standing for several important applications, namely those related to ADL.

65

66 A major group of previous work in activity recognition includes knowledge  
67 and logic-based approaches [23, 16]. For example, authors in [35] proposed  
68 a monitoring system for analysis and recognition of human activities. It in-  
69 cludes detecting, tracking people and recognising some pre-defined activities  
70 using posture information. Three sources of knowledge were exploited: the  
71 model of activities, the 3D model of the observed scene, and the 3D model  
72 of the mobile object present in the observed scene. In [7], a knowledge-based  
73 method is proposed for older people monitoring. Events are modelled as a  
74 function of human body context (e.g., sitting, standing, walking), that is  
75 obtained from images, and the environment context, which is obtained from  
76 accelerometers attached to objects of daily living (e.g., TV remote control or  
77 doors use). A rule-based reasoning engine is used for processing, analysing

78 both context types and detect events that fit in rules. While logic-based  
79 approach is a natural way of incorporating domain knowledge, for every  
80 deployment it requires an extensive enumeration by a domain expert. In  
81 addition, there are some methods that utilise Markov logic networks (MLN)  
82 to model events using first-order logic in a Markov network [18, 8]. In [18],  
83 they represent each target activity as weighted and undirected trees, starting  
84 from primitive actions at the bottom to activities at the top. In [8], an MLN  
85 is constructed to recognise ADL in a smarthome using non-visual and non-  
86 wearable sensors. To overcome the noisy and unreliable observations coming  
87 from the sensors, they build logical models can be checked by human and  
88 linked to domain knowledge.

89

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



97 [9] introduces the Switching Hidden Semi-Markov Model (S-HSMM), a two-  
98 layered extension of the hidden semi-Markov model (HSMM) for modelling  
99 low-level and high-level temporal structures of activities. They show that the  
100 proposed S-HSMM performs better than the HSMMs and the HMMs in the  
101 recognition of frequent and infrequent activities. A recent trajectory-based  
102 approach for human activity recognition [14] combines hierarchical Dirichlet  
103 process and HMM to address some limitations of HMM, especially in pre-  
104 dicting the number of human motion states in videos. But it requires a lot  
105 of computation to obtain the number of motion states. The HMM-based  
106 approaches tries to recognise activities by modelling the time-series func-  
107 tion of events and learning the parameters of the function using supervised  
108 learning techniques. However, recognising complex events, such as “prepar-  
109 ing meal”, using time sequence is very difficult since the sequential pattern  
110 is person-dependent. The unsupervised methods include works such as [17]  
111 in which authors learn motion patterns in traffic surveillance videos by us-  
112 ing a two-layered trajectory clustering in space and time via fuzzy k-means  
113 algorithm. This idea has been extended in [24] and a three-layered cluster-  
114 ing is performed on trajectories in order to learn the variations in spatial  
115 routes, time duration and speed. Then, the spatio-temporal dynamics of

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

129

130 In the literature, there are some methods that use hierarchical models for  
131 activity recognition [19, 1, 34]. The described system in [34], extracts fea-  
132 tures from wearable sensor data and use a two-layered Bayesian network to  
133 model the relation between sub-activities and activities. The sub-activities  
134 and conditional probabilities are learned from data but the activities are

135 manually specified. A method that uses passive sensors in smart home en-  
136 vironment and a two-layered HMM to model relation between sub-activities  
137 and activities is proposed in [19]. Similarly, the system learns sub-activities  
138 from data by clustering. For high-level activities, a HMM is trained using  
139 manually labelled data. In [1], using the trajectories extracted from a fixed  
140 camera, a human behavioural analysis system is proposed. Using time delay  
141 neural networks, first, trajectories are classified into four groups: walking,  
142 running,loitering and stopping. Then, a rule-based fuzzy system is used to  
143 infer macro and group behaviours. The disadvantages of this system is that  
144 training is required for neural networks and the fuzzy system requires specific  
145 rules to recognise activities, which is not an easy task for complex activities.  
146 On the contrary, in our method, without the need of manually annotated  
147 ground truth, we automatically learn the hierarchical relations between ac-  
148 tivities and sub-activities in an unsupervised way.

149

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

### 151 **3 Overview of the proposed Activity Discov-** 152 **ery framework**

153 The complete framework that we proposed in this paper can recognise long-  
154 term activities (hours) in an unsupervised manner and can be used in un-  
155 structured scenes. In order to build a hierarchical activity model that char-  
156 acterizes a complex activity, it uses contextual information to create auto-  
157 matically an intermediate structure of a basic activity. This is performed  
158 by following steps: (i) long-term videos are processed in order to obtain  
159 important information (features) about an observed person (i.e. global po-  
160 sitions and the motion of his/her body parts), (ii) features are used to learn  
161 the multi-resolution levels of the scene regions (topology), (iii) features and  
162 scene regions are combined together to build primitive events which repre-  
163 sent a primitive state transitions within regions, (iv) based on the primitive  
164 events, activities are discovered and the model of an activity is built, (v) the  
165 recognition is performed by comparing similarity between models of activity.

## 166 4 Low-level video processing and primitive 167 events

### 168 4.1 Low-level video processing

169 Our low-level processing is based on two phases: extracting Perceptual Fea-  
170 ture Chunks and learning Topologies.

#### 171 4.1.1 Perceptual Feature Chunks

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

177

178 The position of a person, is estimated using a set of tracklets which is com-  
179 puted for each video chunk by tracking particular corner points. First, 500  
180 corner points [30] are randomly initialized and tracked over time using KLT  
181 [5]. Second, we compute 4 clusters (k-means) of the points with respect to  
182 their speed and position, representing static, slow, medium and fast motion.

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

185

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

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

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

191

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

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

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

198

199 Consequently, each video chunk is associated with a PFC that includes fol-  
200 lowing attributes :  $Departure_{PFC}$ ,  $Arrival_{PFC}$  which are two Gaussian dis-  
201 tributions characterizing the position of the person at the beginning and the  
202 end of the video chunk. The mean and standard deviation  $(\mu, \sigma)$  of the po-  
203 sition distributions are computed using the first (or last)  $n_g$  points of the  
204 global trajectory.  $StartFrame_{PFC}$ ,  $EndFrame_{PFC}$  represent the first and  
205 last frame number of the video chunk, respectively.  $PixelTracklets_{PFC}$  are  
206 the pixel-based tracklets used to calculate the global trajectory of the person.  
207 An example of  $PixelTracklets_{PFC}$  (pink to purple) of a person moving from  
208 the armchair to the kitchen is represented in Figure 1-(a). An illustration of  
209 the PFC attributes are presented in Figure 1-(b). The feature chunks enable  
210 to collect the necessary information for activity understanding and to avoid  
211 expensive computational time, especially for long-term activities. The repre-  
212 sentation contains minimal but important information about the activity in  
213 the scene. For instance, we can store the trajectory information of a 4-hour  
214 video in less than 14Kb of memory.

### 215 4.1.2 The Topology

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

221

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

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

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

231



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

## 238 4.2 Primitive Events

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

242

243 *Primitive Events* are the events characterizing Perceptual Feature Chunks  
244 (section 4.1.1) over a single topology (section 4.1.2). For each person, a se-  
245 quence of *Primitive Events* is built using the sequence of PFCs and a topol-  
246 ogy  $T_l$ . In practice, we build 3 sequences of *Primitive Events* (for  $l = 1, 2$   
247 and 3) for a single video.

248

249 *Primitive Events* has 2 attributes, called  $Transition_{PE}$  and  $LocalDynamics_{PE}$ ,

250 that contain extracted features and their semantic interpretation.

#### 251 **4.2.1 The $Transition_{PE}$**

252 It describes the movement of a person over the scene by extracting the transi-  
253 tion information performed between learned scene regions  $SR_i^l$  at one level,  $l$ .

254

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

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

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

#### 258 **4.2.2 The $LocalDynamics_{PE}$**

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

265

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

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

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

## 278 5 Building the Hierarchical Activity Model

### 279 5.1 The process of Activity Discovery

280 The sequences of  $Primitive Events$  are very informative about the activity  
 281 occurring in the video. However, a  $Primitive Event$  can only describe a

282 snapshot of the person motion. In order to provide more meaning, a better  
 283 representation of the discovered activity is needed.

284

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

- 290 • The *Stay* pattern characterizes an activity occurring within a single  
 291 topology region like "at.region.P", and it is defined as a maximal sub-  
 292 sequence of *Primitive Events* with the same *Transition<sub>PE</sub>*:

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

- 293 • The *Change* pattern describes the transition of the person between re-  
 294 gions like "changing.from.P.to.Q" which is composed of a single *Primitive Event*:

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

295 We define a discovered activity (DA) at a level  $l$  as an extracted *Stay<sub>P-P</sub>* or  
 296 *Change<sub>P-Q</sub>* pattern:

$$DA_{P-Q}^l = Stay_{P-P} | Change_{P-Q} \quad (8)$$

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

302

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

## 309 **5.2 The Hierarchical Activity Model**

### 310 **5.2.1 Definition of the model**

311 We represent the model of an activity as a tree of nodes that is obtained  
312 by merging the set of  $\{DA_s^{l=1,2,3}\}$  ( $s$  is the semantics of the activity) and  
313 has a hierarchical structure based on the three levels of granularity (i.e.  
314  $\{N^{l=1}, \{N_i^{l=2}\}_{1 \leq i \leq n}, \{N_j^{l=3}\}_{1 \leq j \leq m}\}$ ). The tree of nodes represents how dif-

315 ferent activities and sub-activities are connected to each other thanks to a  
 316 set of *attributes* and *sub – attributes* obtained from the properties such as  
 317 type, duration, etc. In other words, a node  $N$  is characterized by *attributes*  
 318 and *sub – attributes*:

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

### 324 5.2.2 Learning phase of the model

325 For a selected instances of the same discovered activities  $DA_s^l$  (e.g.  $s =$  “cook-  
 326 ing”), we learn the model of activity by constructing a tree of nodes where  
 327 each node of level  $l$  is built from the set of discovered activities that are at the  
 328 same resolution level  $l$ ,  $\{DA_{s_1}^l, DA_{s_2}^l, \dots, DA_{s_n}^l\}$  where  $s_1, s_2, \dots, s_n$  are parts  
 329 of  $s$  (i.e. sub-activities of cooking). An example of the constructing pro-  
 330 cess of a tree of nodes from three sequences of discovered activities classified  
 331 from the coarser to the finer one is illustrated in Figure 5-(a). We construct  
 332 an independent model for each type of discovered activity. In the following

333 subsections, we describe the parameters of *attributes* and *sub – attributes*.

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

336 • *Type*: it is adopted from the DAs composing a node. For a node  $N$ ,

337 
$$type_N = type_{DA_s}$$

338 • *Instances*: the amount of training instance of activities composing a  
339 node.

340 • *Duration*: a Gaussian distribution  $N(\mu_d, \sigma_d^2)$  describing the temporal  
341 duration of the training instances.

342 • *Histogram of Local Dynamics  $H(\theta)$*  : is a histogram that charac-  
343 terizes the length and the angle of local motion. As it is presented in  
344 Figure 5-(b), the length is the magnitude of the local motion vector  
345 and the angle is orientation of the vector with respect to x-axis, which  
346 is discretised into 8 bins.

347 **The sub-attribute of a node** The sub-attributes enable us to get infor-  
348 mation from the child nodes. To compute the sub-attributes of a node, we use  
349 the attributes of its child nodes. For a node  $N^l$ , we define two sub-attributes

350 named  $mixture_{sub-activity}$  and  $timelapse_{sub-activity}$  which aim at describing  
 351 two properties of the child nodes  $N_i^{l+1}$  of  $N^l$ :

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

- 355 •  $K$ , is the total number of components (Gaussians) and equal to  
 356 the number of unique  $Types$
- 357 •  $O$ , is the total number of discovered activities at level  $l$  ( $DA^l$ ).
- 358 •  $w_{q=1\dots K}$ , is the prior probability of the component  $q$ . It is equiv-  
 359 alent to the weight of each Gaussian in the MOG. It is computed  
 360 based on the number of appearances of the nodes with the same  
 361  $Type$ :

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

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

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

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



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

### 369 5.2.3 Recognition phase of the model

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

375

376 First, a new sequence of Perceptual Feature Chunks are computed for the  
377 new video. Second, using three levels of topology learned from training  
378 videos, we create new *Primitive Events*. Thereby,  $Transition_{PE}$  of new  
379 *Primitive Events* are matched with the  $Transition_{PE}$  of *Primitive Events*  
380 used in training. Third, the activity discovery process is performed with  
381 the new *Primitive Events* and a new sequence of discovered activities are  
382 computed. Fourth, for each type of discovered activity of the new video, an  
383 activity model is built as explained in Section 5.2.2. Finally, we compute a

384 score between the new model and learned models and classify the activity by  
385 assigning the label of the best match.

386

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

## 395 **6 Experimental results**

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

400

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

407

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

414

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

419

420 The classification results for HOMECARE and HOSPITAL datasets are  
 421 based on leave-one-out cross validation. The evaluation is performed by  
 422 learning the scene (Section 4.1.2) and activity (Section 5.2.2) models from the  
 423 training videos and by recognising activities in a test video. In the HOME-  
 424 CARE dataset, the scene and activity models are learned after processing  
 425 6 videos. The remaining video from the HOMECARE dataset is used for  
 426 recognition procedure. First, activities are discovered in the remaining video  
 427 using the set of extracted *Primitive Events* associated to the person and  
 428 the scene model learned from 6 videos. Then, for each discovered activity,  
 429 an activity model is created and compared with the activity model learned  
 430 from training videos (Section 5.2.3). Similarly, in the HOSPITAL dataset the  
 431 scene and activity models are learned using 3 videos and one video is selected  
 432 to recognise activities. For the CHU dataset, we have randomly selected 10  
 433 videos for learning the scene and activity models. The remaining videos are  
 434 used to recognise activities.

435 To evaluate the framework, we have used True Positive ( $TP$ ), False Posi-  
 436 tive ( $FP$ ), False Negative ( $FN$ ) and calculated *Sensitivity* and *Precision*  
 437 as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

438

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

439

440 An example of learned scene model for the HOMECARE dataset is repre-  
 441 sented in Figure 2. The performance of our framework for HOMECARE,  
 442 HOSPITAL and CHU datasets are displayed in Table 1-(a), Table 1-(b) and  
 443 Table 1-(c), respectively. The recognition results of the proposed framework  
 444 are obtained by comparing with manually annotated ground truths. It can  
 445 be seen from Table 1 that the proposed method proved to be very reliable  
 446 for activities between two areas (e.g. “armchair” to “table” in Table 1-(a),  
 447 ”exercise 1” in Table 1-(b) and “office desk” to “drugs desk” in Table 1-(c)).  
 448 The proposed method is also good at recognising activities occurring in one  
 449 area (e.g. “reading in the armchair” in Table 1-(a), “preparing coffee” in  
 450 Table 1-(b) and “preparing drugs” in Table 1-(c)). Thanks to the proposed  
 451 hierarchical model of activities (section 5.2), we are able to recognise differ-  
 452 ent activities performed within a particular area. For example, the Table  
 453 1-(a) shows that the system has detected two different activities (standing  
 454 and reading) when the person is in the armchair area ( $2^{nd}$  area for  $k = 5$  in  
 455 Figure 2-(b)). The discovery and distinction between two different activities

456 occurring in the same area are possible thanks to the use of the local dy-  
457 namics (section 4.2.2). This can also be seen in Table 1-(b) where the two  
458 different exercises (Up/down and Balance) occurring in the same area are  
459 not confused.

460

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

473

474 Considering the results in Table 1, it can be seen that the framework achieves

475 a high rate of True Positive and a low rate of False Negative. In total, ma-  
476 jority of the performed activities are recognised by the framework.

477

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

486

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

495 In Table 3, we have compared the proposed framework with our previous  
496 work described in [28] and a rule-based method proposed in [35] where the  
497 activities are manually modelled by setting rules and constraints. In [35],  
498 they cannot differentiate finer activities inside an area (e.g, “sitting at the  
499 table” and “eating“) and recognition performance for some activities are not  
500 presented. Thus, for the method in [35], we have given only the results they  
501 have presented and the accumulated recognition rate for merged activities.  
502 The bold values in the table show the best result for each activity class. It  
503 can be seen that for all activity classes the proposed method gives a better  
504 rate of sensitivity and precision compared to the method in [35]. Unlike in  
505 [35], it can be seen that the HAM is capable of differentiating finer activities.  
506 Compared to the method in [28], the proposed HAM enables us to enhance  
507 the recognition results. In two activities (“eating“ and ”preparing meal“) we  
508 achieve better sensitivity and precision rates and in two activities (“inside  
509 bathroom“ and ”from armchair to table“) we achieve better precision rates.



## 510 **7 Conclusion**

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

522

523 The contributions of the framework are twofold: primitive events and hi-  
524 erarchical activity models. To bridge the semantic gap we have proposed  
525 an intermediate layer of primitive events which are used to link semantics  
526 with perceptual information. Thanks to this intermediate layer, the proposed  
527 method overcomes the problem of manually describing the target activities.  
528 The hierarchical activity model give us the opportunity to categorize complex

529 activities using increasing granularity levels of the spatio-temporal structure  
530 of basic activities.

531

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

544

545 The framework can also be used in many other fields such as video surveil-  
546 lance of metros and airports. Our future work is going to be the extension  
547 of our framework to detect abnormal activities in such applications. In addi-

548 tion, we are going to test our framework in online-learning mode by updating  
549 (or creating) the zones and activity models in time with new trajectory in-  
550 formation.

## 551 References

- 552 [1] G. Acampora, P. Foggia, A. Saggese, and M. Vento. Combining neural  
553 networks and fuzzy systems for human behavior understanding. In *IEEE*  
554 *Ninth International Conference on Advanced Video and Signal-Based*  
555 *Surveillance (AVSS)*, pages 88–93, 2012.
- 556 [2] JK Aggarwal and Michael S Ryoo. Human activity analysis: A review.  
557 *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.
- 558 [3] AJ Bharucha, C Atkeson, D Chen, et al. Caremedia: Automated video  
559 and sensor analysis for geriatric care. In *Annual Meeting of the American*  
560 *Association for Geriatric Psychiatry*, 2006.
- 561 [4] Aaron F. Bobick and Andrew D. Wilson. A state-based approach to  
562 the representation and recognition of gesture. *IEEE Transactions on*  
563 *Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, 1997.
- 564 [5] Jean-Yves Bouguet. Pyramidal implementation of the affine lucas  
565 kanade feature tracker description of the algorithm. *Intel Corporation*,  
566 2001.

- 567 [6] Simone Calderara, Rita Cucchiara, and Andrea Prati. Detection of ab-  
568 normal behaviors using a mixture of von mises distributions. In *IEEE*  
569 *Conference on Advanced Video and Signal Based Surveillance (AVSS)*,  
570 pages 141–146. IEEE, 2007.
- 571 [7] Yuanyuan Cao, Linmi Tao, and Guangyou Xu. An event-driven context  
572 model in elderly health monitoring. In *Symposia and Workshops on*  
573 *Ubiquitous, Autonomic and Trusted Computing, UIC-ATC '09.*, pages  
574 120–124, July 2009.
- 575 [8] Pedro Chahuara, Anthony Fleury, Francois Portet, and Michel Vacher.  
576 Using markov logic network for on-line activity recognition from non-  
577 visual home automation sensors. In *Ambient Intelligence*, volume 7683  
578 of *Lecture Notes in Computer Science*, pages 177–192. Springer Berlin  
579 Heidelberg, 2012.
- 580 [9] T.V. Duong, H.H. Bui, D.Q. Phung, and S. Venkatesh. Activity recogni-  
581 tion and abnormality detection with the switching hidden semi-markov  
582 model. In *IEEE Computer Society Conference on Computer Vision and*  
583 *Pattern Recognition (CVPR)*, volume 1, pages 838–845 vol. 1, 2005.

- 584 [10] Alireza Fathi, Jessica K Hodgins, and James M Rehg. Social interac-  
585 tions: A first-person perspective. In *IEEE Conference on Computer*  
586 *Vision and Pattern Recognition (CVPR)*, pages 1226–1233. IEEE, 2012.
- 587 [11] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily  
588 actions using gaze. In *Proceedings of the 12th European Conference on*  
589 *Computer Vision - Volume Part I, ECCV'12*, pages 314–327, Berlin,  
590 Heidelberg, 2012. Springer-Verlag.
- 591 [12] Anthony Fleury, Michel Vacher, and Norbert Noury. Svm-based multi-  
592 modal classification of activities of daily living in health smart homes:  
593 sensors, algorithms, and first experimental results. *IEEE Transactions*  
594 *on Information Technology in Biomedicine*, 14(2):274–283, 2010.
- 595 [13] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, and  
596 Deva Ramanan. Computational studies of human motion: part 1, track-  
597 ing and motion synthesis. *Found. Trends. Comput. Graph. Vis.*, 1(2-  
598 3):77–254, 2005.
- 599 [14] Qingbin Gao and Shiliang Sun. Trajectory-based human activity recog-  
600 nition with hierarchical dirichlet process hidden markov models. In *IEEE*

- 601        *China Summit International Conference on Signal and Information Pro-*  
602        *cessing (ChinaSIP)*, pages 456–460, July 2013.
- 603 [15] Shaogang Gong and Tao Xiang. Recognition of group activities using  
604        dynamic probabilistic networks. In *Computer Vision, 2003. Proceedings.*  
605        *Ninth IEEE International Conference on*, pages 742–749. IEEE, 2003.
- 606 [16] Somboon Hongeng, Ram Nevatia, and Francois Bremond. Video-based  
607        event recognition: activity representation and probabilistic recognition  
608        methods. *Computer Vision and Image Understanding*, 96(2):129–162,  
609        2004.
- 610 [17] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, Dan Xie, Tieniu Tan, and  
611        Steve Maybank. A system for learning statistical motion patterns. *IEEE*  
612        *Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1450–  
613        1464, 2006.
- 614 [18] Gowun Jeong and H.S. Yang. Context-aware activity recognition by  
615        markov logic networks of trained weights. In *16th International Con-*  
616        *ference on Virtual Systems and Multimedia (VSMM), 2010*, pages 5–12,  
617        Oct 2010.

- 618 [19] Tim L.M. Kasteren, Gwenn Englebienne, and Ben J.A. Krose. Hier-  
619 archical activity recognition using automatically clustered actions. In  
620 *Ambient Intelligence*, volume 7040 of *Lecture Notes in Computer Sci-*  
621 *ence*, pages 82–91. Springer Berlin Heidelberg, 2011.
- 622 [20] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Pro-*  
623 *ceedings of Ninth IEEE International Conference on Computer Vision*  
624 *(ICCV)*, pages 432–439. IEEE, 2003.
- 625 [21] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozen-  
626 feld. Learning realistic human actions from movies. In *IEEE Conference*  
627 *on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE,  
628 2008.
- 629 [22] James MacQueen et al. Some methods for classification and analysis of  
630 multivariate observations. In *Proceedings of the fifth Berkeley symposium*  
631 *on mathematical statistics and probability*, volume 1, page 14. California,  
632 USA, 1967.
- 633 [23] Gérard Medioni, Isaac Cohen, François Brémond, Somboon Hongeng,  
634 and Ramakant Nevatia. Event detection and analysis from video



- 635 streams. *IEEE Transactions on Pattern Analysis and Machine Intel-*  
636 *ligence*, 23(8):873–889, 2001.
- 637 [24] B.T. Morris and M.M. Trivedi. Trajectory learning for activity under-  
638 standing: Unsupervised, multilevel, and long-term adaptive approach.  
639 *Pattern Analysis and Machine Intelligence, IEEE Transactions on*,  
640 33(11):2287–2301, Nov 2011.
- 641 [25] Matthai Philipose, Kenneth P Fishkin, Mike Perkowitz, Donald J Pat-  
642 terson, Dieter Fox, Henry Kautz, and Dirk Hahnel. Inferring activities  
643 from interactions with objects. *Pervasive Computing, IEEE*, 3(4):50–57,  
644 2004.
- 645 [26] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily  
646 living in first-person camera views. In *IEEE Conference on Computer*  
647 *Vision and Pattern Recognition (CVPR)*, pages 2847–2854. IEEE, 2012.
- 648 [27] Fatih Porikli. Learning object trajectory patterns by spectral clustering.  
649 In *IEEE International Conference on Multimedia and Expo (ICME'04)*,  
650 volume 2, pages 1171–1174. IEEE, 2004.
- 651 [28] Guido Pusiol, Francois Bremond, and Monique Thonnat. Unsupervised  
652 discovery, modeling, and analysis of long term activities. In *Computer*

- 653        *Vision Systems*, volume 6962 of *Lecture Notes in Computer Science*,  
654        pages 101–111. Springer Berlin Heidelberg, 2011.
- 655 [29] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing hu-  
656        man actions: a local svm approach. In *Proceedings of the 17th Inter-*  
657        *national Conference on Pattern Recognition (ICPR)*, volume 3, pages  
658        32–36. IEEE, 2004.
- 659 [30] Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Com-*  
660        *puter Society Conference on Computer Vision and Pattern Recognition*  
661        *(CVPR'94)*., pages 593–600. IEEE, 1994.
- 662 [31] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath  
663        Gupta, and Ramesh Jain. Content-based image retrieval at the end of  
664        the early years. *IEEE Transactions on Pattern Analysis and Machine*  
665        *Intelligence*, 22(12):1349–1380, 2000.
- 666 [32] Ekaterina H. Spriggs, Fernando De La Torre, and Martial Hebert. Tem-  
667        poral segmentation and activity classification from first-person sensing.  
668        In *IEEE Computer Society Conference on Computer Vision and Pattern*  
669        *Recognition Workshops (CVPR Workshops)*, pages 17–24. IEEE, 2009.

- 670 [33] Raghav Subbarao and Peter Meer. Nonlinear mean shift over riemannian  
671 manifolds. *International Journal of Computer Vision*, 84(1):1–20, 2009.
- 672 [34] A. Subramanya, A. Raj, J. Bilmes, and D. Fox. Hierarchical models  
673 for activity recognition. In *IEEE 8th Workshop on Multimedia Signal*  
674 *Processing*, pages 233–237, 2006.
- 675 [35] Nadia Zouba, Francois Bremond, and Monique Thonnat. An activity  
676 monitoring system for real elderly at home: Validation study. In *Seventh*  
677 *IEEE International Conference on Advanced Video and Signal Based*  
678 *Surveillance (AVSS)*, pages 278–285. IEEE, 2010.

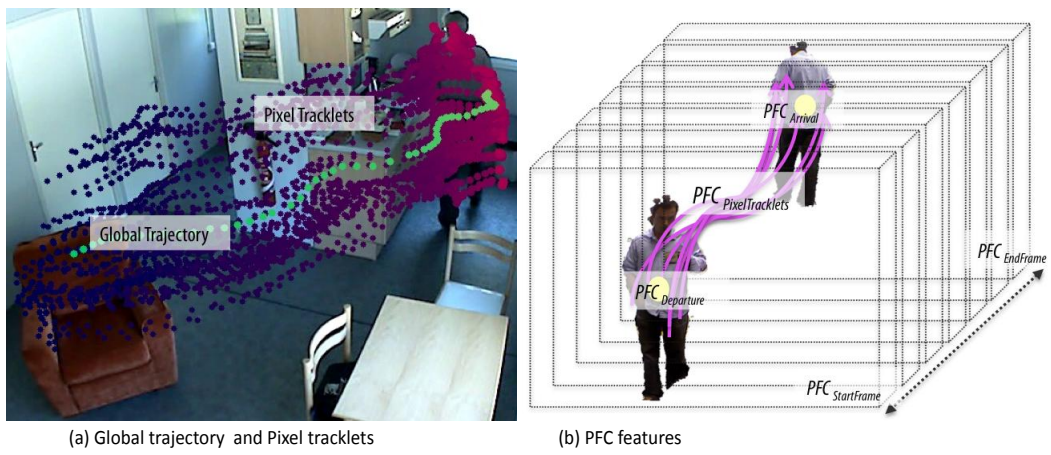


Figure 1: Global trajectories (green) and Pixel Tracklets (purple to pink) to construct Perceptual Feature Chunks.



(a) HOMECARE

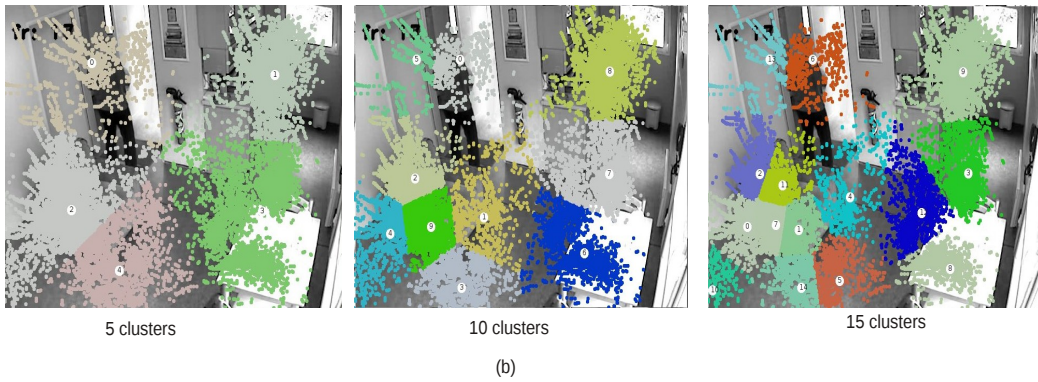


Figure 2: (a) The empty scene for HOMECARE. (b) Example of the scene model with  $l = 1, 2$  and  $3$  obtained by k-means clustering ( $k = 5, 10$  and  $15$ ) for HOMECARE dataset described in section 6.

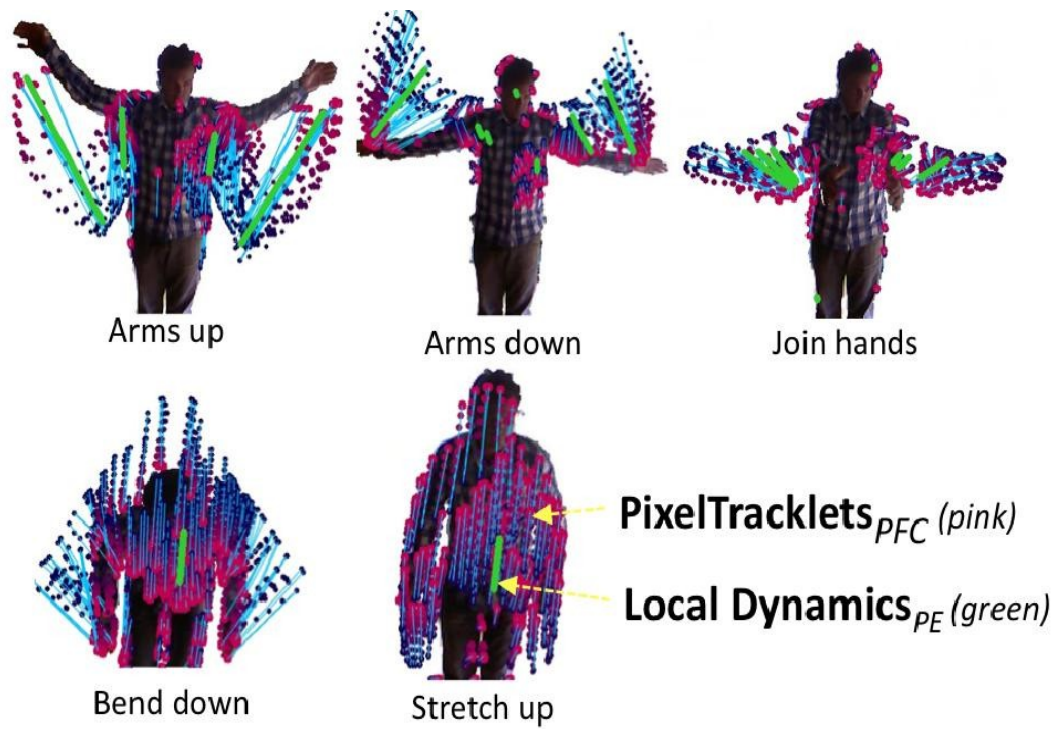


Figure 3: Example of the abstraction of  $PixelTracklets_{PFC}$  (pink) into  $LocalDynamics_{PE}$  (green). Each,  $LocalDynamics_{PE}$  is displayed as a straight line corresponding to the start and end points of an abstracted tracklet (blue line).

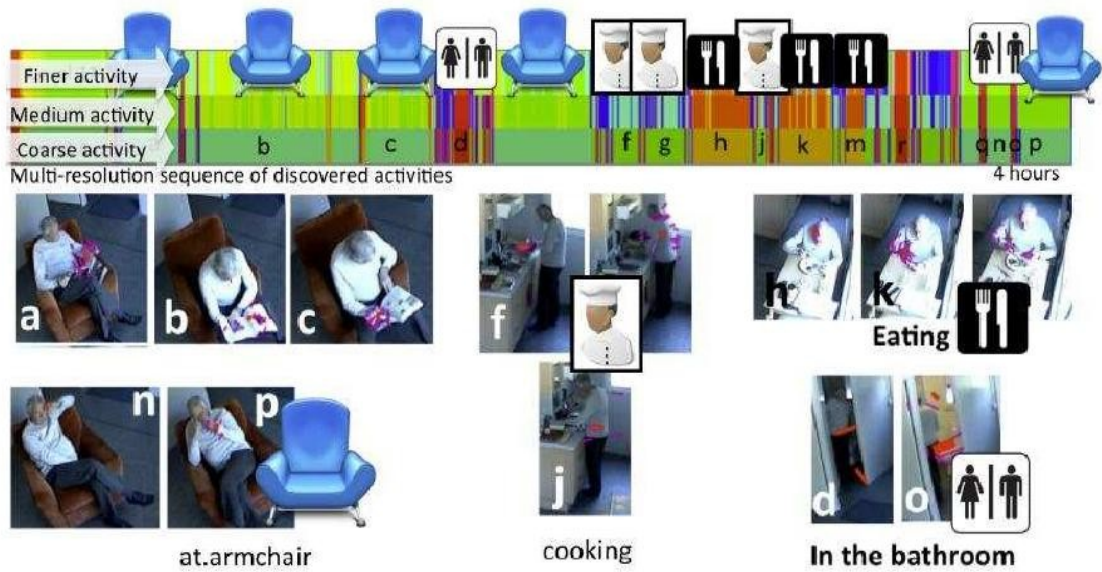
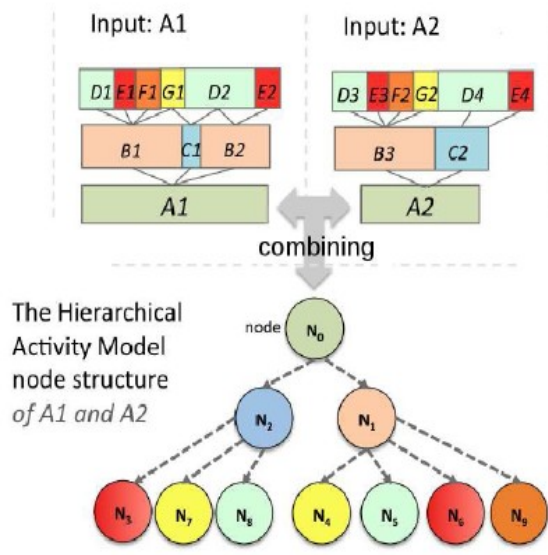
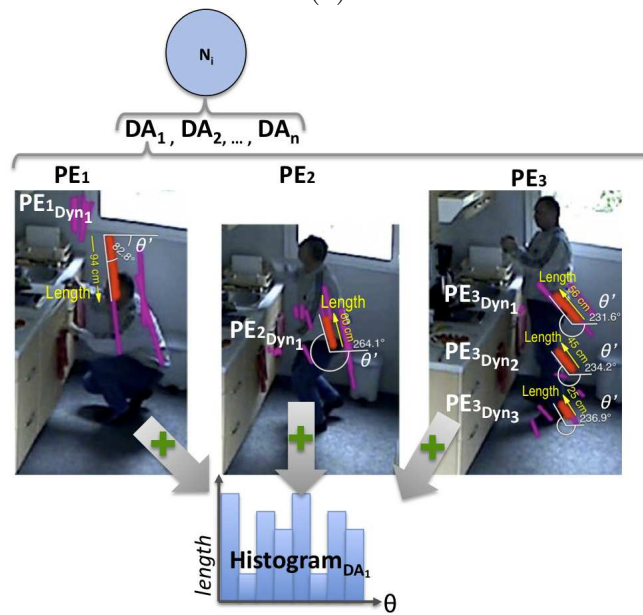


Figure 4: Example of discovered activities (coloured segments) for 4 hours video of one person performing everyday activities. 5 actions in the armchair (a, b, c, n and p), 3 cooking (f, g and j) and eating (h, k and m) actions and 2 actions in the bathroom (d and o) are discovered.



(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.



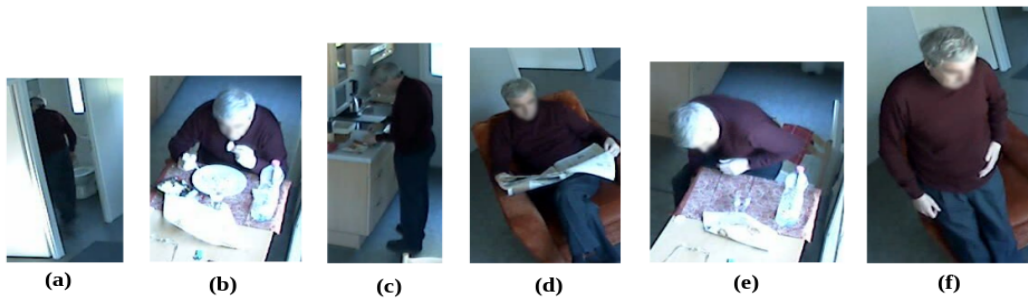


Figure 6: Everyday activities in HOMECARE dataset: (a) In the bathroom, (b) Eating, (c) Preparing meal, (d) Reading in the armchair, (e) Sitting at eating place, (f) Standing at armchair.

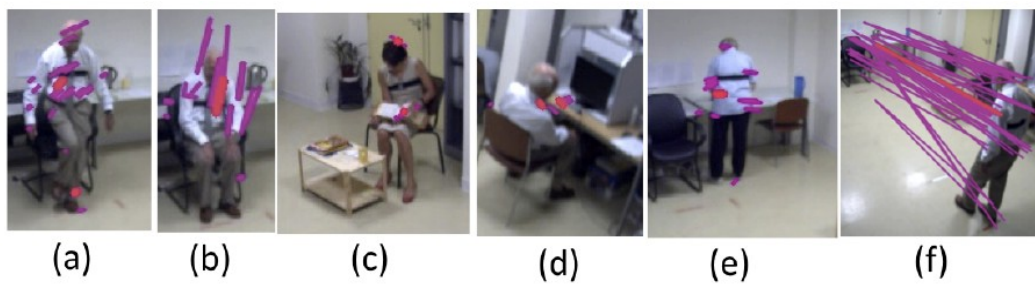


Figure 7: Guided activities in HOSPITAL dataset (a) Balance: Standing on one foot at at time, (b) Up/Down: Standing and sitting down in a continuous way, (c) Reading at the table, (d) At the computer, (e) Preparing coffee, (f) Exercise1/Exercise2: moving from the chair to a marked position and coming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

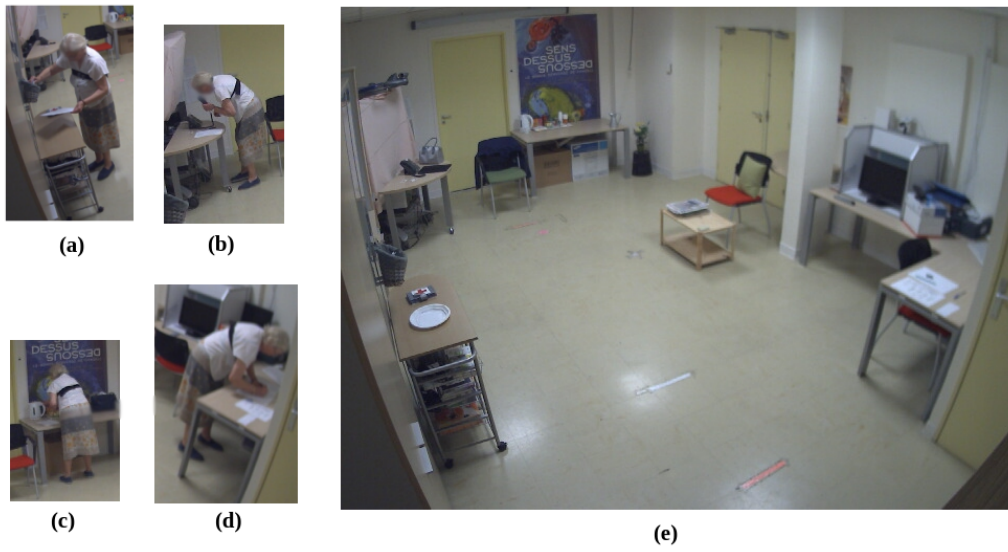


Figure 8: Non-guided activities in CHU dataset: (a) Preparing drugs, (b) Talking on the phone, (c) Preparing tea, (d) Paying bill. (e) Overview of the hospital room in HOSPITAL and CHU datasets.

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	<b>100</b>	<b>96</b>
Sitting at eating place			<b>88.88</b>	<b>100</b>	<b>88.88</b>	<b>100</b>
Reading in armchair	85.96	80.32	<b>100</b>	<b>85.71</b>	<b>100</b>	82.75
Preparing meal	80	57.14	88	85	<b>98.18</b>	<b>96.42</b>
Standing at arm chair	-	-	<b>100</b>	<b>84.61</b>	<b>100</b>	<b>84.61</b>
Inside the bathroom	-	-	<b>100</b>	77.78	<b>100</b>	<b>87.5</b>
From armchair to table	-	-	<b>100</b>	88.89	<b>100</b>	<b>94.11</b>
From armchair to kitchen	-	-	<b>100</b>	<b>93.75</b>	<b>100</b>	<b>93.75</b>