Expert Systems with Applications 39 (2012) 10873–10888

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living

Alexandros André Chaaraoui, Pau Climent-Pérez, Francisco Flórez-Revuelta *

Department of Computing Technology, University of Alicante, P.O. Box 99, E-03080 Alicante, Spain

ARTICLE INFO

Keywords:
Human behaviour
Ambient-Assisted Living
Computer vision
Motion analysis
Action recognition
Activity recognition
Activities of daily living (ADLs)

ABSTRACT

Human Behaviour Analysis (HBA) is more and more being of interest for computer vision and artificial intelligence researchers. Its main application areas, like Video Surveillance and Ambient-Assisted Living (AAL), have been in great demand in recent years. This paper provides a review on HBA for AAL and ageing in place purposes focusing specially on vision techniques. First, a clearly defined taxonomy is presented in order to classify the reviewed works, which are consequently presented following a bottom-up abstraction and complexity order. At the motion level, pose and gaze estimation as well as basic human movement recognition are covered. Next, the mainly used action and activity recognition approaches are presented with examples of recent research works. Increasing the degree of semantics and the time interval involved in the HBA, finally the behaviour level is reached. Furthermore, useful tools and datasets are analysed in order to provide help for initiating projects.

1. Introduction

Human Behaviour Analysis—and Understanding—(HBA, HBU) involves a wide range of investigation fields from motion detection and background extraction to expert systems and high-level abstraction behaviour models. This paper targets two purposes: On the one hand, researchers need to categorise existing works assuming a common taxonomy and a clear differentiation basis. On the other hand, as the application areas of these fields grow constantly; stable areas, like Video Surveillance, are covered thoroughly; while other more recent areas, like Ambient-Assisted Living (AAL) and ageing in place at smart home scenarios, present a lack of unifying works and recent state-of-the-art reviews. This makes initiation in these areas difficult, also because of the involvement of a wide variety of pure research areas from artificial intelligence to natural language processing.

For this reason, this paper deals with the state-of-the-art of HBA/HBU from an Ambient Intelligence (AmI) point of view, focusing especially on indoor scenarios and techniques which are designed for AAL purposes. This way, recognition of activities of daily living (ADLs) covers the main interest of this paper. Nevertheless, it is necessary to first face a classification of HBA levels, and to deal with all the necessary previous tasks.

To avoid the common difficulties present in vision-based systems (such as occlusions, view-dependent features, lightning conditions, etc.), occasionally these systems are enhanced with other sensors; mostly binary sensors and RFID labels. Therefore, although vision will be focused on mainly, other complementary sensors involved will be discussed briefly too.

The remainder of this paper is organised as follows: Section 2 goes through taxonomies which are applied by other authors and presents an abstraction, degree of semantics and time-oriented classification which is used in the rest of the paper. Section 3 deals with the lowest level, i.e. pose, gaze and motion estimation. These elements are used as *action primitives* in Section 4 where human actions are recognised based on video data and other sensor data fusion (RFID tags, accelerometers, etc.). Section 5 focuses on activity recognition methods which are of special interest in AAL: ADLs in indoor environments, like cooking and grooming, are recognised with different approaches detailed in that section. Finally, Section 6 deals with behaviour recognition methods that establish the highest degree of abstraction. Section 7 summarises some of the most used datasets and tools in the reviewed works that are available.

2. HBA taxonomies

In this section, different Human Behaviour Analysis taxonomies from some of the most recent and relevant research works are discussed in order to point out differences and converge at a well-defined classification of the works analysed in following sections.

Moeslund, Hilton, and Krüger (2006) defined an action taxonomy which has been adopted in later works and subsequent surveys. From lower to higher degree of abstraction three levels are defined:

* Corresponding author. Tel.: +34 965903681; fax: +34 965909643.
E-mail addresses: alexandros@dtic.ua.es (A.A. Chaaraoui), pcliment@dtic.ua.es (P. Climent-Pérez), florez@dtic.ua.es (F. Flórez-Revuelta).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.005

- Basic motion recognition derives in so called *action* or *motor primitives* representing the atomic entities out of which actions are built. Therefore, as stated in Poppe (2010), an action primitive is an atomic movement that can be described at the limb level.
- A set of different or repetitive *action primitives* make up an *action*.
- Involving a larger scale of events, the context of the environment and the interacting objects or humans it is possible to recognise the actual *activity*.

This way, when making a cup of tea, single movements of arms and hands would be *action primitives*; placing the kettle on the stove or grabbing a cup from the cupboard would be *actions*; and finally, the whole process would make up an *activity* as different actions and interaction with several objects are involved.

Although this taxonomy is clearly defined and quite often referenced in HBA-related papers, most researchers use their own taxonomy, as usefulness depends on research goals and application areas. Since this classification is particularly focused on actions, it is difficult to adapt to higher level approaches, where the main targets are ADLs and behaviour analysis.

In Wu, Osuntogun, Choudhury, Philipose, and Rehg (2007), *activities* are defined as the combination of *actions* and *objects*. Whereas actions are recognised by a set of *verbs*, objects or places are recognised by a set of *nouns* which are targets of actions. Instead of recognising the actions, object recognition is tackled in order to infer human activities. Turaga, Chellappa, Subrahmanian, and Udrea (2008) distinguish between actions and activities by defining that activities involve coordinated actions among a small number of humans.

Regarding behaviour analysis, Ji, Liu, Li, and Brown (2008) define behaviours as human motion patterns involving high-level description of actions and interactions. In contrast to Moeslund et al. (2006), dependence on the context of the environment, objects and human interaction are taken into account at the behaviour level. In Monekosso and Remagnino (2010), behaviours are understood as patterns in a sequence of observations of activities or events. Activities such as *cooking*, *eating*, *watching TV* or *no detectable activities*; and events from the environment, emitted by binary sensors installed in smart homes, enable to recognise repeatable patterns and detect anomalies.

In this paper, HBA tasks are classified into *motion, action, activity* or *behaviour* levels regarding the degree of semantics and the amount of time involved in the analysis. Therefore, Fig. 1 shows that both the time frame taken into account and the degree of semantics (DoS) involved in the recognition and classification process grow as we reach a higher level of the pyramid.

At the *motion* level, tasks such as movement detection, and background extraction and segmentation are faced (Hu, Tan, Wang, & Maybank, 2004; Moeslund et al., 2006; Porle, Chekima, Wong, & Sainarayanan, 2009). Using a time frame in units of frames, a lot of research is done in the field of gaze and head-pose estimation (Launila & Sullivan, 2010; Ozturk, Yamasaki, & Aizawa, 2009; Reale, Hung, & Yin, 2010, 2010; Rybok, Voit, Ekenel, & Stiefelhagen, 2010; Shimizu & Poggio, 2003).

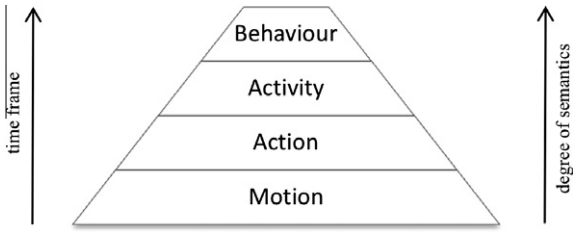

Fig. 1. Human behaviour analysis tasks – classification.

At the *action* level, human motion is not only detected, but also recognised in order to establish what a person is doing or with which objects the person is interacting. In a time frame in units of seconds, simple human activities; like sitting, standing or walking (Bao & Intille, 2004; Chung & Liu, 2008; Liu, Chung, & Chung, 2010; Lester, Choudhury, & Borriello, 2006; Zhou et al., 2008); can be recognised; as well as location changes in indoor and outdoor environments (Nait-Charif & McKenna, 2004).

At the *activity* level, a set of multiple actions is classified in order to understand human behaviour in a time frame from tens of seconds to units of minutes. ADLs are recognised; like *cooking*, *taking a shower* or *making the bed*; as those require tracking and classification of a sequence of actions in a particular order. This way, the sets of actions are understood as activities, where these activities are either the goals or the results of their involving human actions.

At the *behaviour* level, highly-semantic comprehension comes into play. Within a time frame ranging from days to weeks; ways of living, personal habits, and timetables and routines of ADLs can be analysed. At this point, abnormal behaviours and anomalies can be detected, for instance, in order to be able to detect senile dementia prematurely (Karaman et al., 2010; Mihailidis, Boger, Craig, & Hoey, 2008; Mihailidis, Carmichael, & Boger, 2004).

Table 1 summarises the different degrees of semantics considered by the taxonomy, along with some examples. Not only time frame and semantic degree grow at higher levels of this hierarchy, but also complexity and computational cost lead to heavy and slow recognition systems, as each level requires most of the previous level tasks to be done too. For this reason, level abstraction is key in order to analyse only the necessary parts and avoid redundant processes. Human tracking is the best example because it can be approached at least at the first three levels, having different tracking targets and using different kinds of features from the underlying levels. Therefore, tracking will not be discussed in this paper on its own, but tracking approaches from the analysed works will be mentioned when significant.

3. Pose, gaze and motion estimation

Motion recognition is the basis for estimation of human pose and gaze direction (also referred to as focus of attention) and for further HBA tasks. Motion can be seen as a series of poses along the time; the human body is an articulated system of rigid segments connected by joints (as models used in Andriluka, Roth, & Schiele (2009) and Sapp, Toshev, & Taskar (2010) assume); and human motion is often considered as a continuous evolution of the spatial configuration of the segments or body posture (as stated in Li, Zhang, & Liu (2010) and exploited in Andriluka et al. (2009) and Sapp et al. (2010)). On the other hand, the gaze can either be seen as a line in the 3D space or a cone; or, if working only in the horizontal plane (as some works do, as seen later on); a direction and an angle.

3.1. Pose estimation

There are handfuls of previous surveys which analyse and describe "human motion" or "human behaviour understanding" (Hu et al., 2004; Jaimes & Sebe, 2007; Poppe, 2007, 2010; Wang, 2003); earlier works, as is logic, review lower level techniques (e.g. the work by Gavrila (1999)); and later works review also further abstraction levels, approaching more to what is understood as *behaviour* by the taxonomy employed in this review.

Table 1
Classification of tasks according to the degree of semantics (DoS) involved.

DoS	Time lapse	Description
Motion	frames, seconds	Movement detection, background subtraction and segmentation; gaze and head-pose estimation
Action	seconds, minutes	Establish with which objects the person is interacting. Recognise simple human primitives (sitting, standing, walking, etc.)
Activity	minutes, hours	Tasks that consist of a sequence of actions in a particular order. ADLs are recognised (e.g. cooking, taking a shower or making the bed)
Behaviour	hours, days, ...	Highly-semantic comprehension comes into play (ways of living, personal habits, routines of ADLs)

Gavrila's work deals with a great amount of techniques which are aimed at providing the area of interest, without an explicit shape model by using either segmentation (as background subtraction or skin detection) or Haar wavelets (and PCA).

When shape models are used, XYT volumes are built (which reveal characteristic patterns, see Fig. 2) (Gorelick, Blank, Shechtman, Irani, & Basri, 2007); stick figure models are also employed. Others use a 'blob finder': each blob is defined as the shirt, pants, hands and head of a person, and these are found in an image. There is also a subsection dedicated to 3D body modelling.

When it comes to action recognition, the document presents a variety of techniques, which can detect actions (at the level defined in the taxonomy employed throughout this document). Most techniques are either based on Dynamic Time Warping (DTW), as well as Hidden Markov Models (HMM).

In the work by Moeslund (2001), bigger emphasis is given to the recognition of motion and actions, whereas activities and behaviours are treated to a minor extent. In a later work by the same author (Moeslund et al., 2006), activity and behaviour recognition are dealt more widely (this work will be mentioned with more detail in further sections).

Nevertheless, the techniques revealed in Moeslund (2001) allow a greater understanding of current methods, and thus, both this and (Gavrila, 1999) can be used to introduce the early phases of behaviour analysis which is widely seen as a post processing, or a step to follow after prior segmentation and low-level analysis.

The work by Moeslund et al. (2006), can be divided into two major parts: pose and motion capture, and action recognition. In the first part, a series of works are introduced and three different phases namely model initialisation, tracking and pose estimation are presented. These three phases are then subdivided into families of approaches due to some similarity; or they are divided into subphases.

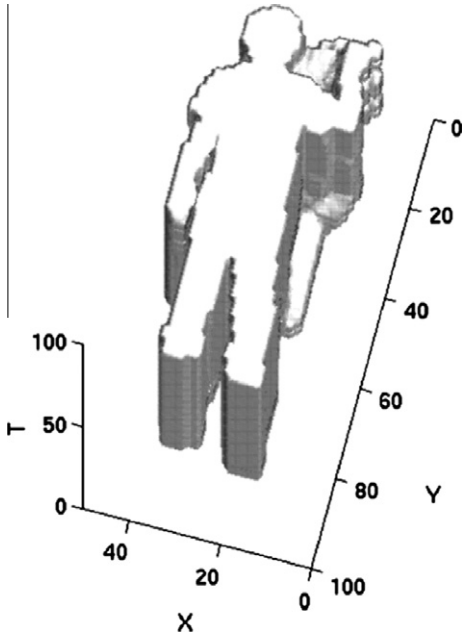

Fig. 2. Gorelick et al. (2007) XYT volume. Reprinted from Turaga et al. (2008).

Model initialisation, which captures prior knowledge of a specific person in order to constrain tracking and pose estimation, is thus presented from different points of view: the kinematic structure being used (e.g. a skeleton with a number of joints with specified Degrees of Freedom, DoFs) and how it is initialised; the technique used to approximate the subject's shape (either using simple shape primitives such as cylinders, cones, etc., or a polygonal mesh); or appearance (presence of skin, clothing, the use of body part detectors for different types of limbs or the trunk, see Fig. 3; similarly to what is described in Andriluka et al. (2009), Bourdev, Maji, Brox, & Malik (2010), Eichner & Ferrari (2009), Ferrari, Marin-Jimenez, & Zisserman (2008), Park & Kautz (2008), Sapp et al. (2010) and Shotton et al. (2011)).

Figure-ground segmentation techniques (tracking) are then introduced according to the approach being used. Six different families are seen: (1) background subtraction; (2) motion-based segmentation; (3) appearance-based segmentation; (4) shape-based segmentation; (5) depth-based segmentation; and (6) temporal correspondences.

Pose estimation techniques are introduced next, and they are classified into three groups as in Moeslund (2001), according to the presence and use of an explicit model: model-free, indirect model use, direct model use.

In what is related to model-free approaches, research about 'body plans' and combinations of body part detectors are presented (Wu & Nevatia, 2005), more modern works include (Andriluka et al., 2009; Bourdev et al., 2010; Eichner & Ferrari, 2009; Ferrari et al., 2008; Sapp et al., 2010). Furthermore, techniques that are example-based are presented too; these use either a representation of the mapping from 2D silhouette sequences in image space to skeletal motion in 3D pose space (Rosales & Sclaroff, 2000) or direct lookup of silhouette sequences for recognition (Agarwal & Triggs, 2004; Howe, 2004; Shakhnarovich, Viola, & Darrell, 2003; Sminchisescu, Kanaujia, Li, & Metaxas, 2005). There are two main drawbacks to this family of techniques: (1) when using 2D silhouettes as key poses, a constraint on the point of view is added, which limits recognition to that exact point of view; to overcome this, different viewpoints can be added into the database, which can be cumbersome and lead to worse inter-class recognition; and (2) the larger the number of classes, the worse the recognition will work (see for instance (Winn, Criminisi, & Minka, 2005)).

In indirect model use techniques, in turn, methods that use direct reconstruction of both model shape and motion from the visual-hull are revealed (Cheung, Baker, & Kanade, 2003; Mikić, Trivedi, Hunter, & Cosman, 2003).

Finally, direct model use techniques are seen; under this set of methods, multiple view 3D pose estimation using gradient descent techniques, and more recently particle filtering (a state space search reduction is used), are of interest. An evolution of the works presented under this section of Moeslund's work, can be seen in the papers by Bandouch, Engstler, and Beetz (2008) and Beetz, Bandouch, Jain, and Tenorth (2010); which are commented later on. To end with direct model use techniques, monocular 3D pose estimation and learnt motion model methods are presented.

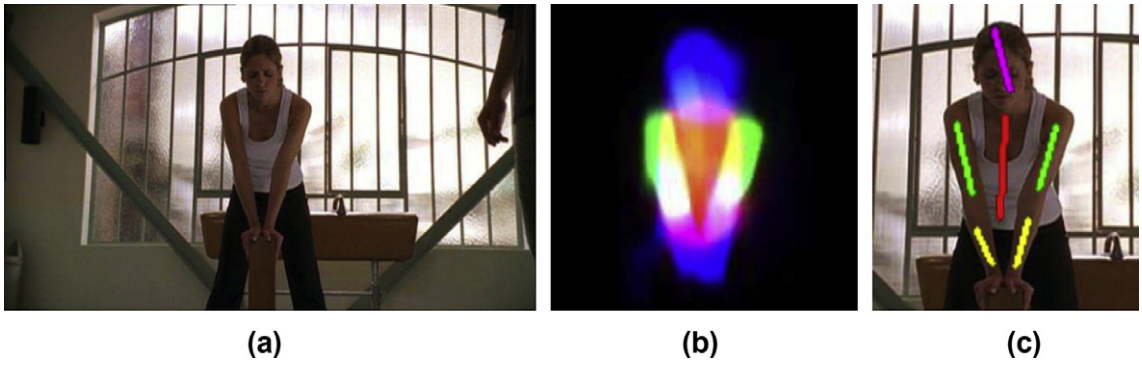

Fig. 3. Upper-body part detector results as stated in Ferrari et al.'s work. Reprinted from Ferrari et al. (2008).

A more recent survey is the one by Hu et al. (2004); this paper is highly comprehensive, as it starts with motion detection and object classification and tracking, but also deals with HBU; it divides the approaches to this task depending on the techniques used (Dynamic Time Warping, Finite State Machines, Hidden Markov Models, Time-Delay Neural Networks, Syntactic Techniques, Non-deterministic Finite Automata, Self-Organising Maps).

Then, it also reveals techniques aimed at describing the behaviours using natural language. After that, it goes onto personal identification by means of different methods including all sorts of biometrics (limb lengths, speeds, gait, height, weight, face recognition, etc.).

It also takes the fusion of data from multiple cameras into account, and explores future developments. In this section, however, only the first phases, which deal with motion detection, are of interest. Here, pose estimation algorithms which prove useful for the tasks described in Section 1, as well as Human–Computer Interaction (HCI), will be revealed.

In the literature, very different approaches (Mcivor, 2000; Toyama, Krumm, Brumitt, & Meyers, 1999) for foreground/background segmentation can be found, which are very popular in computer vision tasks. Such techniques are aimed at determining the position of the moving objects in a scene. Frame differencing might be the earliest of these techniques, followed by other background subtraction techniques, based on a wide variety of models. Statistical models could be dealt here, such as Adaptive Background Mixture Models (using mixture of Gaussians) (Stauffer & Grimson, 1999), Wallflower (Toyama et al., 1999), or others (Mcivor, 2000; Toyama et al., 1999).

Segmentation algorithms, if not accompanied by further techniques, provide only very basic pose estimation (as a silhouette or area), which can be only used as information on *where* the subject is. Depending on the task to be done, this could be enough; otherwise, further estimation refinement techniques are at hand. For example, Lv and Nevatia (2007) take a single silhouette from a video, and with it, they are able to recognise actions from a set of previously learnt examples.

Other ways to determine the position of objects in a scene are based on object detection, without segmentation. In the work by Dalal et al. (2005), descriptors are generated for image windows. These provide a means for recognition of certain shapes (the human body). These descriptors are then fed to a classifier for training. After that, in the test stage, the classifier determines whether each sample in a window is of the object class. Applying this method, object presence in a scene can be confirmed. It can also be applied to determine the kind of object present in the bounding box obtained from a previously applied segmentation algorithm.

Segmentation-free pose estimation techniques are based on human models (or object models in general); either complex anthropomorphic 2D or 3D models (i.e. van der Meulen & Seidl, 2007, used in Bandouch et al. (2008)), or approximations to it (cylinders or ovals (van der Meulen & Seidl, 2007), skeletons and stick figures (Carranza, Theobalt, Magnor, & Seidel, 2003), etc.). The main difficulty that arises when using no segmentation, is how to find a way to determine where the joints of people are (Tao, Hu, & Zhou, 2007).

Nevertheless, those models can also be used with segmentation; with that coarse knowledge extracted from the previous phase where silhouettes are estimated, recent works propose to infer finer pose estimation. To do so, Boulay, Bremond, and Thonnat (2006) propose a system which, using a single camera, is able to determine the pose from a set of poses. Using the data from the segmentation phase, they proceed to estimate 3D cues, such as depth and others. With these, and using a virtual 3D environment with the same size and camera position, they calculate how the person might appear as a 2D silhouette depending on the pose in those exact coordinates. After that, the virtually generated poses and the actual pose are compared with a histogram comparison algorithm, which yields the most probable pose as a result. The whole workflow runs in real time (4–15 fps). Further work (Zouba, Boulay, Bremond, & Thonnat, 2008), applies the described technique in an AAL scenario. This way, there are works which choose to cover the blob with an elliptical model in order to extract conclusions based on the direction of the major and minor axis of the ellipse and the length ratio between them (see Nait-Charif & McKenna, 2004). For instance, this method enables to detect if the individual is standing, sitting or lying. As a result, the system is able to detect possible falls.

A 3D-volume approach is presented by Anderson et al. (2009), which uses 2D silhouettes extracted from calibrated cameras to mount what is called a 'person voxel'. They further apply fuzzy logic to classify three different body poses (upright, in-between and on-the-ground); and thus be able to detect falls.

More recently, other works based on similar approaches have appeared (Bandouch et al., 2008; Beetz et al., 2010); these allow a very detailed pose estimation. These are based on the use of three cameras, instead of one, without prior knowledge of the room model (no calibration), the only requisite being that the cameras are set so that they see the object from different perspectives (ideally orthogonally, with non-overlapping views). With such a scheme, the *MeMoMan* project (Bandouch et al., 2008) has been able to determine very fine pose estimation in real time. The technique is similar to Boulay et al. (2006) and Zouba et al. (2008). Using the silhouettes (in this case three), a hierarchical 3D human model is applied, which *fits* into the observed silhouettes. To reduce the dimensionality, mathematical models are used, which allow faster pose estimation. Furthermore, the system has been applied in a kitchen scenario, for ADL analysis (Beetz et al., 2010).

Fig. 4. Focus of attention estimation of a group of people. Reprinted from Canton-Ferrer et al. (2008).

3.2. Focus of attention/gaze estimation

So far, only works related to body pose estimation have been discussed. In the field of AAL, other body cues are interesting too, such as gaze direction (or focus of attention), which provides further information related to what is being done in the scene (Marin-Jimenez, Zisserman, & Ferrari, 2011). Gaze estimation can also be used to detect distraction, or abnormal situations. On the other hand, different researchers treat gaze differently, it can be understood as a line in the 3D space (a beam-like approach) or a cone (Canton-Ferrer, Segura, Pardas, Casas, & Hernando, 2008); or, if working only in the horizontal plane (Launila & Sullivan, 2010; Ozturk et al., 2009), a direction and an angle (compass-like approach). Some of the reviewed works use gaze estimation for different purposes. Canton-Ferrer et al. (2008), for instance, use gaze estimation for attention analysis in smart classrooms or offices (see Fig. 4). Doshi and Trivedi (2010a, 2010b) use gaze estimation to detect driving styles and distractions.

In Marin-Jimenez et al. (2011), gaze estimation is seen as an additional cue for video annotation and understanding. The authors present various techniques with different degrees of fineness; by either using simple information about *left* or *right* head orientation, or yaw and pitch angles along with inferred depth (z-axis) information. For the upper-body and head detectors described, the model of (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010) is used.

Head pose estimation can be understood as an alternative term for the detection of the focus of attention of a specific person, as the gaze of a person is determined to some extent by the pose of his/her head (Marin-Jimenez et al., 2011). In Launila and Sullivan (2010), different colour and shape properties are combined in order to estimate the pose of the players' heads in a soccer match. This work is part of a greater project whose final objective is to reconstruct a whole soccer match in 3D and in real-time. This would make it possible to watch the match from any point of view and solve conflictive situations at refereeing.

Within the field of head pose estimation, in Ozturk et al. (2009), gaze direction is obtained in steps of 22.5° in wide indoor areas such as airports and malls. They develop the first solution to the head pose estimation problem using only the data proceeding of a single 2D camera. A two level particle filter made up of colour and edge histograms is used for tracking. Afterwards the individual silhouette is matched to one of 16 patterns by using shape descriptors and SIFT points.

4. Action recognition

After the initial step of motion detection, and pose or gaze estimation; basic actions can be understood as a series of motions detected; either in the whole or in some parts of the subject's body (arms, legs, head, etc.); or in some areas of a room (living room, kitchen, bathroom, outdoors, etc.). Such actions can be recognised because of the different body poses which are involved, and the variation through short periods of time.

In order to understand the difference between an action and an activity, not only time lapse is taken into account; the objects and people involved are important too. For instance, a person manipulating an object is performing an *action* (say, opening a lid); while several of such *actions*, performed with different objects, compose what is called an *activity* (e.g. *cooking a meal*).

This section is centred in the former, and as such, it will describe the works and techniques which are either based on or aimed at recognising *actions*. Various categorisations have been found for action recognition methods (Li et al., 2010; Turaga et al., 2008; Weinland, Boyer, & Ronfard, 2007). These works classify action recognition based on the approaches used for action modelling.

According to Turaga et al. (2008), approaches for modelling actions can be categorised into three major classes: nonparametric, volumetric, and parametric time-series approaches (see Fig. 5).

Inside 2D-template-based approaches described in Turaga et al. (2008), temporal templates are estimated for each action class. The work by Bobick and Davis (2001) presents two mechanisms called Motion-Energy Image (MEI) and Motion-History Image (MHI), respectively. The MEI is a binary image which represents where motion has occurred, while the MHI is a scalar-valued image where intensity of the pixels is a function of the recentness of motion (Fig. 6).

Also under this class, Ben-Arie, Wang, and Pandit (2002) reveal a technique, in which data representing the angle of the limbs is extracted from the silhouettes. Only a few representative poses are taken into account for the learning process and stored in a database. Indexing is performed, so that searches relating only one specific limb can be performed over the action model database. For recognition, after extracting each limb's pose, a voting scheme is used, which returns the most likely action.

In volumetric approaches, in contrast, features are not extracted on a frame-by-frame basis, but instead the whole video is considered as a 3D volume of pixel intensities, and standard image features are extended to deal with the 3D case (Turaga et al., 2008). For instance, Laptev (2005) proposed an spatio-temporal (ST) generalisation of the Harris interest point detector, to model and recognise actions in space–time. Niebles, Wang, and Fei-Fei (2008) use such interest points in a bag-of-words model in order to represent actions. Clustering of the features and classifiers (such as SVMs, graphical models, etc.) can be applied afterwards. Based on similar ideas, Ryoo and Aggarwal (2009), propose a kernel function to measure similarity between pairs of videos.

Similarly to Turaga et al. (2008), Weinland et al. (2007) categorise the works in action recognition depending on the way they model the actions being recognised. The two approaches mentioned are 'model-based' or 'template-based' ('holistic'). According to the authors, the former approach assumes a known parametric model, typically a kinematic model, and actions are represented in a joint or parameter space. This approach shows difficulties to estimate the pose correctly without the use of markers or other means of easing the task.

In contrast to this, in the latter approach, the use of spatio-temporal shapes as action templates reduces the burden. Actions are then modelled using information retrieved from the images (such as silhouettes, optical flow, etc.). For recognition, comparison

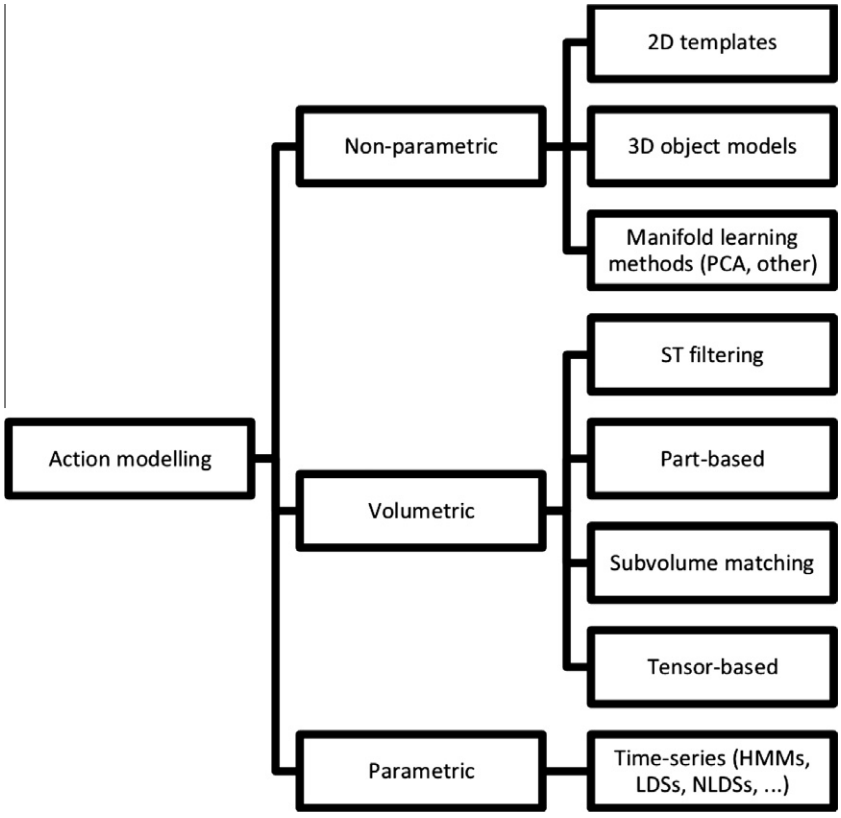

Fig. 5. Turaga et al.'s (2008) action modelling classification.

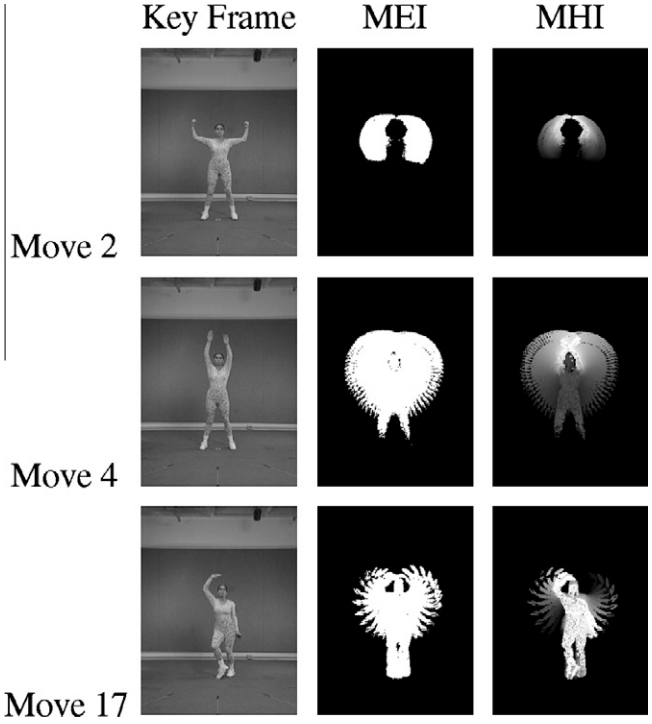

Fig. 6. Examples of MEI and MHI images. reprinted from Bobick and Davis (2001).

between the observation-inferred templates and learnt templates is required. The authors point out the limitation of this latter technique: both learning and recognition need to be done under similar camera configurations.

In Li et al. (2010), a similar categorisation is made. Here, the approaches are either based on the extraction of action descriptors from the silhouette sequences (similarly to 'model-based' approaches seen in Weinland et al. (2007) or the non-parametric methods shown in Turaga et al. (2008)); or based on the extraction of features from each silhouette and model the dynamics of the action explicitly (again, similarly to the aforementioned 'holistic' approaches).

The approach proposed by the authors in Weinland et al. (2007) takes advantage of the template based methods, but without this last restriction, thus they avoid the need for view-specific training databases. Action recognition is performed from 2D cues, while in the learning stage, the HMM states are not represented by a single 3D exemplar; this accounts for different body proportions, style, or clothing. In Weinland et al. (2007, Section 2), a state of the art relating view-independent action recognition can also be found.

Another application of action recognition is that of smart surveillance systems. Cheng, Yang, Han, and Sawhney (2008), for instance, propose a system which is used both for parking lot surveillance and indoor tracking in which primitive actions are logged. The technique they present is based on Histogram Oriented Occurrences (HO2), which is described as "(...) a new feature that captures the interactions of all entities of interest in terms of configurations over space and time. HO2 features encapsulate entity tracks, inter-object relationships and the context of the environment into a spatial distribution that characterises the corresponding event." These new features allow easier multi-agent event recognition; in contrast to simpler feature vectors which yield HMMs with too many nodes. The proposed feature is based on the ideas of Histograms of Oriented Gradients (HOGs), presented in Dalal et al. (2005) and Lowe (2004) and seen in the previous section; and the Shape Context descriptor, described in Belongie, Malik, and Puzicha (2000). After the HO2 features are calculated,

they are then fed to an SVM classifier, in order to detect the different event types (arrivals, departures, trunk loading or unloading, etc.).

Cherla, Kulkarni, Kale, and Ramasubramanian (2008) use DTW along with a two-component feature vector, whose elements are: the width profile, calculated from the silhouette's bounding box; and some spatiotemporal features, such as the displacement of the centroid (in X and Y), and the standard deviation (also in both axes). Because of the dimensionality reduction applied, actions performed sidewise (e.g. *walking*) can be detected better than actions which take place in a frontal manner (except for *hand waving* and other actions which involve sidewise movements of the limbs).

Three works by Oikonomopoulos, Pantic, and Patras (2008), Oikonomopoulos, Patras, and Pantic (2005) and Oikonomopoulos, Patras, and Pantic (2009) describe techniques aimed at recognising basic actions. The earlier work (Oikonomopoulos et al., 2005) reveals a technique for aerobic exercises action recognition by means of a sparse representation based on spatio-temporal salient points; these are obtained using a method presented by Kadir and Brady (2003) that takes scale variance into account, among other considerations. A distance measure between two points is also introduced, which is based upon the chamfer distance.

In Oikonomopoulos et al. (2008), the authors deal with the use of *B-splines* as a means of describing the movement of points along the time axis; these descriptors will feed a codebook afterwards, which allows further recognition to be performed. The later paper (Oikonomopoulos et al., 2009) is based on ST shape model. The goal is to detect actions without prior segmentation. In the paper, a ST segmentation is performed, and a voting scheme is used. The technique is robust against partial occlusions, and multiple activities can be detected if performed simultaneously.

Within the Prometheus (FP7) project, a work by Quintas, Khoshhal, Aliakbarpour, Hofmann, and Dias (2011) describes the use of Concurrent Hidden Markov Models (CHMMs), for the detection of ADLs in a smart home environment. Although the work talks about behaviours, the events described along the paper correspond to what is classified as actions by our taxonomy. Thus, this research falls into the category treated in the current section.

4.1. Data fusion

Other works base their motion recognition techniques in other kinds of sensing, which are more intrusive than the methods described up to this point; these are based either on other kinds of vision, or on other kinds of sensing.

One remarkable example is 'wearable vision', which is a semi-intrusive scheme, in which a camera is mounted on the person's shoulder, or attached to the frame of their glasses, etc. Papers reviewed using such techniques (Ren & Philipose, 2009; Sun, Klank, & Beetz, 2009; Sundaram & Cuevas, 2009) emphasise how wearable cameras avoid the body to occlude what is being managed with the hands (see Fig. 7). They also point it as a more natural approach because activities are performed looking at what is being done. These head-mounted cameras, allow the researchers to work with 'first-person' images, in which they see the hands of the user, and the object interaction along their ADLs. The only drawback of such a scheme is that hands provoke occlusions too, just as the body does in non-wearable cameras. Moreover, cameras of such kind need to be improved, as they can be really cumbersome for the final users.

Other sensing devices are also used in diverse works (Altun & Barshan, 2010; Bao & Intille, 2004; Maurer, Smailagic, Siewiorek, & Deisher, 2006; Spriggs, De La Torre, & Hebert, 2009; Yang et al., 2008), or a combination of vision and RFID tags (Wu et al., 2007) either as a direct way to recognise the objects being manipulated,

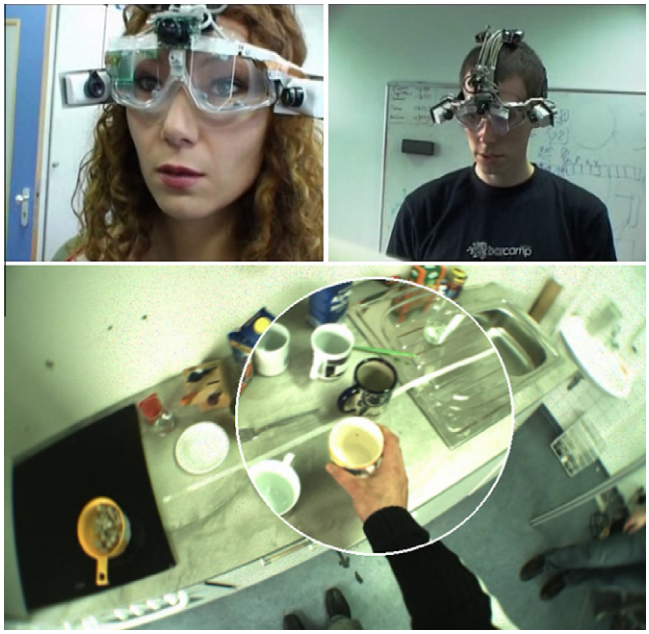

Fig. 7. A gaze directed camera used in Sun et al. (2009); the camera itself is shown in the upper images. The lower image shows a superimposed camera view generated by the device (reprinted).

or as a means for supervised learning of visual object appearance (using the correspondence between the RFID and visual data). Yang et al. (2008) use a sensor network in order to determine the position of the body by means of a network of wireless motion sensors. In Spriggs et al. (2009), data from the wearable camera is enriched with data from the Inertial Measurement Units (IMUs), which are worn by the subjects in a form of bracelets. Furthermore, Maurer et al. (2006) recognise activities by using only one bracelet (called *eWatch*) which they attach at different body parts for comparison. Bao and Intille (2004) use bi-axial accelerometers attached to different body parts in order to determine the wearer's ADLs (mostly actions, but some complex activities too). Altun and Barshan (2010) also use Inertial/Magnetic Sensor Units (IMSUs) in order to determine the actions being performed. This last work also emphasises the fact that vision and IMSUs are not exclusive. It also mentions some other papers in which vision and the mentioned sensors are used in hybrid systems (Tao et al., 2007), or as a method to check the correct classification when using only data from IMSUs, as in Aminian et al. (1999), Najafi et al. (2003) and Roetenberg, Slycke, and Veltink (2007). Kwapisz, Weiss, and Moore (2011) propose the use of tri-axial accelerometer-equipped smart phones for the same purposes.

Li et al. (2010) take the use of other visual sensors into account. Their technique, based on 3D point clouds, deals with the recognition of actions by means of bags of three-dimensional points, obtained at 15 fps by a depth camera that acquires the depth through structured infra-red light. The proposed method obtains a highly reduced subset of the point cloud, based on the observation that pixels in the silhouette boundary (contours) are the most relevant ones, as they carry more information on the observed body's shape. Action recognition is then performed by means of action graphs, which are described in a previous work by the same authors (Li, Zhang, & Liu, 2008).

5. At activity level: activities of daily living

At this level, the goal of the recognition is to classify a sequence of actions into their targeting *activity*. For instance, some children

could be moving their legs and their arms, jump and run, and interact with a ball. But, *what* are they playing? At this moment, semantics come into play and understanding the compound of actions is what gives actual value to the recognition. How can we distinguish a basketball game from a volleyball game? This is the reason why the particular order of the actions and the interacting objects are key for activity recognition. If we are able to distinguish the type of ball, the net from the hoop, or even that in basketball, the player runs while bouncing the ball by tracking the involved actions; we are taking into account a larger time frame and a significant higher degree of semantics than in previous HBA levels.

In particular, recognising ADLs in smart homes can lead to understand what a person is doing, and enables monitoring of completeness and correctness of these activities. In this matter, Mihailidis et al. (2004) are able to track the activity of hand washing to assist older adults with dementia. Multiple orders in the process can be correct, but not of all of them; their system is able to prompt the user if a necessary step is missing or the order of the implied actions is unacceptable. Vision is used as the only sensor in the developed system for two purposes: (1) tracking of hand location; and (2) tracking of step-specific object locations. In previous works (Mihailidis, Fernie, & Cleghorn, 2000), hand washing was tracked with switches and motion sensors. This way, the system could infer whether the hands were in the sink or if soap was used. Nevertheless, the authors explain that although this data was reliable, too little was known about the user and the environment. Even if taps are on and the motion sensor indicates that the hands are in the sink, there is no guarantee that the individual is actually washing his/her hands.

Related to this type of activity recognition, Wu et al. (2007) stand out in activity recognition based on object use. As mentioned before, these authors define activities as combinations of actions and objects and intend to recognise and track object use in order to infer human activities. Object models are acquired automatically from video, whereas object identification is based on RFID labels. At the learning phase the user wears a RFID bracelet which reads the RFID tags attached to the surrounding objects in a home environment. Assuming that the object being moved is always the object in use and that only one object is being moved at a time, the system learns the relationship between the segmented image and the active RFID tag using a dynamic Bayesian network. As arms and hands move with the objects, skin filtering is applied beforehand. At the test phase, the system works without the RFID data as objects are recognised by detecting SIFT features within the segmented area. These key points are matched based on maximum likelihood to the previously trained SIFT points. As the number of possible SIFT features is very high; clusterization is applied, using the *K*-means algorithm, in order to obtain a delimited histogram of SIFT features for each object.

In Zhou et al. (2008), activity recognition is approached differently. The individual silhouette is obtained at different positions of a living room. Grouped into 10–20 prototypes each silhouette stores its centre, width and height and is manually labelled with a location. A fuzzy inference method is used to estimate the most likely physical location of test silhouettes. Location estimation and previously assigned coordinates enable average speed measurement, which is used besides location in order to recognise human indoor activities. A Hierarchical Action Decision Tree (HADT) is used to classify human actions using multiple levels. At the first level, human actions are classified based on location and speed. With *K*-means, clustering feature patterns are obtained; and activities of daily living, like walking or visiting the bathroom, are recognised; this is achieved by using the *K*-Nearest Neighbour (KNN) method. At the second level, a more precise recognition of activities like washing, eating or cooking is achieved. From the individual silhouette, the smoothed boundary of the human body is extracted using a snake model, which is represented with Hu Moment Invariants (HMI) (Xu & Prince, 1998). This way, the temporal variation of the HMI values are used to measure the level of body motion and instead of tracking single actions, activities are inferred based on how active the person is. The third and last level is only used when a person remains at the same physical location while he/she is moving constantly; this happens, for instance, at exercising. In this condition, further recognition is needed, and primitive visual features are used. By partitioning the video frames in blocks of 8×8 pixels, motion is analysed individually at each block. This way, for a 640×480 pixels video frame, 4800 points are taken into account to characterise low-level motion. Locally Linear Embedding (LLE) is applied to reduce the dimensionality of these feature vectors into a small set of composite features. These features are handled as a trajectory and matched using distance correlation and KNN classification.

When dealing with different types of sensors in smart homes, uncertainty of sensor data needs to be considered. In Hong et al. (2009), belief in sensor data is deeply analysed. When dealing with binary sensors; like movement detectors, contact switch sensors and pressure mats; the sensor data could be erroneous due to a variety of reasons. The sensor itself could be faulty; the data could be approximate, as the exact value is impossible to be measured because of the very nature of what is being measured; or the system could have corrupted the data while reading and sending it to the upper level. Dempster–Shafer's theory of evidence (Dempster, 1968; Shafer, 1976) is considered to be able to represent ignorance due to lack of information, and to aggregate belief when new evidence is obtained. Kitchen door sensors and motions sensors are used to recognise ADLs, like making a drink (differentiating between hot and cold) or making breakfast (cereals, toast or eggs). With multivalued mapping, rules are built in order to know which elements are involved in which actions. For instance, making a tea implies a tea bag necessarily, but milk can be optional. This way, evidence is assigned to these rules, and it is possible to infer the most likely activity given the certainty of the current sensor data.

Nicolini, Lepri, Teso, and Passerini (2010) collected data from a couple who lived at a custom built condominium for a period of 10 weeks. Several hundreds of sensors, including audio–visual recording, collected data in order to be analysed in intervals of 30 s with 15 s of overlapping between each consecutive interval. ADLs like watching TV, grooming, reading and using the phone are recognised with multi-labelled prediction. Besides sensor data, average activity duration is used to train SVM classifiers, one for each activity. As a refinement stage, Conditional Random Fields (CRF) (Sutton & McCallum, 2007) are applied to model sequential observations and recognise completed activities among combinations of local on-going activities. Validation was done using leave-one-day-out cross validation and area under the ROC curve (AUC) as a figure of merit reaching a result from 81% to 97%.

So far, we have presented vision-based activity recognition systems which use global features, like image foregrounds or individual silhouettes; or local features, i.e. keypoints, as salient points or corners. The field of Image Analysis and Processing provides a wide range of image features and types of key points; these are used in diverse application areas and present different advantages and drawbacks (Juan & Gwon, 2009; Tuytelaars & Mikolajczyk, 2007). Although most popular key points techniques as SIFT and SURF are applied frequently at activity recognition, concrete keypoint-based features for this purpose are available too. Velocity history of tracked key points is used in Messing, Pal, and Kautz (2009). Interest points are chosen based on gradient difference and tracked using a KLT tracker (Lucas, 1981). This way, about 500 features are tracked at a time, replacing missed points on the fly; their velocity history is used as the basic feature. Classification is based on a generative mixture model and experimentation data is presented on

the KTH dataset (see Section 7.2), as well as on an own dataset where activities like *writing a phone number on a whiteboard* or *peeling a banana* are recognised (shown in Fig. 8).

In Lester et al. (2006), Lester et al. proposed a personal activity recognition system for health-care purposes which satisfies three key restrictions: (1) data is collected from a single body sensor and it is not required to be from the same point for every user, (2) personalisation could enhance results but should not be required, and (3) should be effective even with a cost-sensitive subset of the used sensors. With these pre-requisites, data of 10 male and 2 female individuals were collected in order to recognise eight different physical activities; like *sitting*, *standing*, *walking* (which we classify as *actions*); but also activities with higher semantics and interactions with objects are recognised, as for instance, *brushing teeth* or *riding an elevator up or down*. Volunteers wore a multi-modal sensor board at wrist, waist or shoulder. This way, data from following sensors were collected simultaneously: microphone, light phototransistor, 3-axis digital accelerometer, 2-axis digital compass, digital barometer/thermometer, digital ambient light (IR and visible) and digital humidity/temperature. Out of these data, 18,000 samples per second are reduced to 651 features; these include linear and log-scale FFT frequency coefficients, cepstral coefficients, spectral entropy, band-pass filter coefficients, correlations, integrals, means and variances. Their recognition algorithm is hybrid in the sense that models of the underlying distributions of the data classes and class boundaries for discriminative techniques are learnt. At a first level, a custom boosting is applied to select the best features iteratively and to feed an ensemble of discriminative static classifiers with the most discriminative subset of features of each activity. At a second level, HMMs are used to recognise activities based on the class probabilities obtained from the static classifiers.

Activities of interest to medical professionals; such as *toileting, bathing* and *grooming*; are recognised in Tapia, Intille, and Larson (2004, chap. 10). Tapia et al. introduce an interesting approach for labelling train samples manually: *the Context-aware Experience Sampling Tool*. Subjects carry a PDA that is used as a timing device to trigger self-reported diary entries. The PDA asks the user for information when a certain number of changing sensor values are detected; multiple choice questions can then be answered by the user. This way, the system is told what activity he/she is performing right now, and for how long he/she has been doing this activity. Tests have been performed for 14 days in two one-bedroom apartments with one subject each. 77 state-change sensors collected data at doors, windows, cabinets, microwave ovens, refrigerators, toilets, showers, water taps, etc. In this work, naïve Bayesian classifiers are extended to incorporate temporal relationships among sensor firings and implemented in two versions: the first is a multi-class naïve classifier, in which the child nodes either have *exist* or *before* attributes. This means that temporal order is being considered: if a particular sensor should fire when performing a certain activity (*exists*), or if it should fire *before* another particular sensor. The second version consists of multiple binary naïve Bayes classifiers, one for each activity. In contrast to the first version, this one is not mutually exclusive. Finally, activity duration is considered too, applying feature windows individually for each activity.

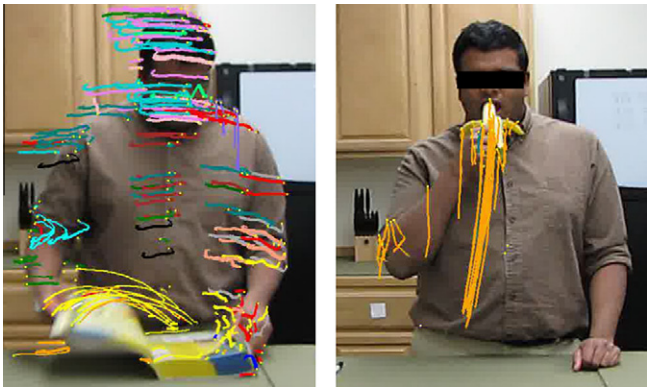

Fig. 8. Labelled example flows from *Lookup in Phonebook* and *Eat Banana*. Reprinted from Messing et al. (2009).

In order to describe ADLs consistently at several abstraction levels, Beetz et al. (2010) proposed so called *Automated Probabilistic models of Everyday Activities* (*AM-EvAs*). AM-EvAs consist of automated activity observation systems, interpretation and abstraction mechanisms for behaviour and activity data, as well as reasoning and query systems that enable AM-EvAs to answer semantic questions about the activities. This way, these models have information about the involved actions and sub-actions, objects, agents who performed the activities, location and time. The purpose of these models is to build a knowledge-based framework to combine observations of human activities with a-priori knowledge about actions, and to make the classification and assessment of actions and situations objective. Therefore, action patterns of different activities are learnt from several subjects in order to support singularities. The activities are observed with vision-based full-body motion tracking, RFID tags and magnetic sensors. The sensor data stream is segmented and classified with action classifiers which recognise movement primitives. This data is combined with time intervals and events, and represented in a first-order logic language. Probability distributions are represented either with Bayesian—or with Markov—Logic Networks. In conclusion, AM-EvAs make it possible to train objective knowledge-based models to save meta-information from activities and query following types of questions: (1) relational knowledge, like *Which is the whole pose sequence of a table setting activity?*, (2) action related concepts, like *Where is the table setting activity performed?*, and (3) probabilistic knowledge, like *Having observed that a bowl has been taken and that an egg has been cracked, how likely is it that brownies are being baked?*

6. Human behaviour understanding

According to the taxonomy being employed, *behaviour* is understood as the highest level of complexity and time span. It is seen as a long-lasting series of activities that *tend to* occur in a certain order. It could be seen as the observed person's daily routines. Deviations from the pattern can be seen as extraordinary, and as such, they give information about the person's evolution (say, health status or independence in the case of elderly people living alone (Monekosso & Remagnino, 2010)).

Under this definition, a number of works using vision as a source for activity detection are found. Although, many of the reviewed works in the field of behaviour analysis and AAL (Cardinaux, Brownsell, Hawley, & Bradley, 2008; Hara, Omori, & Ueno, 2002; Hayes, Pavel, & Kaye, 2008; Jain, Cook, & Jakkula, 2006; Mahmoud, Akhlaghinia, Lotfi, & Langensiepen, 2011; Monekosso & Remagnino, 2010; Park, Lin, Metsis, Le, & Makedon, 2010; Virone & Sixsmith, 2008; Wood et al., 2008) are based on other sensor devices: lights and use of appliances (Cardinaux et al., 2008; Jain et al., 2006; Monekosso & Remagnino, 2010); pressure mats (Cardinaux et al., 2008); basic motion detectors—such as door sensors or similar (Hayes et al., 2008; Park et al., 2010; Wood et al., 2008); infrared sensors—SMDs (Cardinaux et al., 2008; Hara et al., 2002; Jain et al., 2006); health monitoring (Wood et al., 2008); etcetera.

Behaviour of people in the scene is seen in some of these works as the circadian activity rhythm (CAR); that is, the evolution of

ADLs throughout the day (Virone & Sixsmith, 2008; Wood et al., 2008). By learning the CAR of a person, either on a weekly or a 5-day basis, the system can recognise abnormalities in the behaviour as deviations from the previously observed routines.

From the works that do indeed include vision sources, context-aware systems (Brdiczka, Crowley, & Reignier, 2009; Chung & Liu, 2008) and video annotation systems (Robertson & Reid, 2006, Robertson, Reid, & Brady, 2006; Yamamoto, Mitomi, Fujiwara, & Sato, 2006) can be found. These kinds of systems can be classified under this section dedicated to 'behaviour understanding', as the main point of these is not limited to determine the activities people are performing in each moment, but to extract and infer further information from the recognised activities.

The work by Brdiczka et al. (2009) deals with the recognition of situations (context). In the cited work, the authors present a system which tracks persons in 3D using cameras, and is able to extract information about the entities' pose (*role*); speed; and interaction with other entities (either these are people, furniture or appliances), according to the distance to them. In this arrangement, people wear headsets, which detect whether they are talking; and microphones are arranged so that ambient noise can be detected. Numeric codes are given to each possible permutation (single versus multiple people, with or without ambient noise, with or without people talking). Using these codes, which fuse all the collected information, different kinds of 'situations' are learnt and recognised by means of left–right HMMs (these situations are *individual work*, *introduction [of various people to each other]*, *aperitif*, *siesta [of one individual]*, *presentation* or *[board] game [among multiple people]*).

Robertson and Reid (2006) and Robertson et al. (2006) present a video annotation technique which is able to extract the commentary of a tennis sequence. To do this; position, velocity, and action descriptions are fused and fed into an spatio-temporal action recogniser, which is in turn fed to an HMM which applies a smoothing process to the output using model-based scene knowledge (which are modelled manually by using knowledge of the rules of the game and spatio-temporal constraints in the movements of players). Their uppermost layer consists of 'behaviour HMMs', which take the output of the smoothing HMM to recognise sets of activities which form a more general behaviour (e.g. *baseline-rally* or *serve-and-volley* play types in tennis). As further work, application to abnormality detection is presented.

Chung and Liu (2008) go further and propose a system which takes low-level information (such as poses, which the authors call 'activities'), and combine it with other two contexts: spatial (*where* does the activity happen), and temporal (*when* does the activity happen and how long does it last). This way, using a Hierarchical Context Hidden Markov Model (HC-HMM), behaviours are learnt and later recognised from vision sources. Their techniques are applied to AAL in the context of a nursery house in which different behaviours are monitored (see Fig. 9). These include sequences of activities such as *sit down and watch TV for a while*, *go to the toilet*, *lie on bed to sleep*, *eat breakfast*, *take a walk*, etc. The normal sequences of activities are learnt from the behaviour of the monitored people, in order to detect abnormality in the expected routines (either in their duration, time of day, or location).

In Chung and Liu (2008), results are compared with the same set of videos, with other methods such as the presented in Duong, Bui, Phung, and Venkatesh (2005) and Nguyen, Phung, Venkatesh, and Bui (2005). Duong et al. (2005), present what is called the Switching Hidden Semi-Markov Model (S-HSMM), which allows activity duration abnormality detection, although the sequence of activities is restricted both in order and in the number of activities (exactly six); the space where activities happen is also restricted, as the room is divided into a discrete number of 'cells' of 1 square meter each, and 'hotspots' are defined according to how some appliances and tools are arranged. Nguyen et al. (2005) present an application of the hierarchical HMM which detects three different activities, namely *have a snack*, *short meal* or *normal meal*, depending on the visited spots in each action sequence.

Other works are concerned about the recognition of behaviours that include more than one person interacting in the scene. Early works in this field include (Oliver, Rosario, & Pentland, 2000), in which the performance of different HMM-based techniques for the recognition of interactions among two people is compared, and it is concluded that CHMMs (coupled HMMs) perform better for the task. The work introduces an interesting agent framework for the synthesis of artificial behaviours that are used for training along with real-life video data. CHMMs were introduced in Brand (1996), as a better approach for interaction modelling. Hongeng and Nevatia (2001) present a hierarchy of events along with a method for interaction modelling and recognition. Modelling is performed by means of action threads (associated to each actor/agent) and temporal constraints. Recognition is achieved by propagating these constraints and likelihoods in a Temporal Logic Network (TLN).

A more recent work from Liu et al. (2010), extends the methods presented in Chung and Liu (2008) for the recognition of behaviours present in pairs or groups of people. By recognising the number of people present in the scene, switching between two different HMM-based approaches is performed by a switching module called 'Switch Control' (SC). Either Individual Duration HMM (IDHMM) or Interaction-Coupled Duration HMM (ICDHMM) are used as a consequence. The methods, as in the previously presented work, are applied in AAL environments (nursing homes).

7. Useful research tools

When reading a survey like this one, most readers are either initiating themselves in the field, or taking up again and looking for what advances other researchers made in the last few years. Therefore, this section is made for those who are going to start a new project and could take advantage of existing tools, models and datasets; in order to build upon, and be able to compare between different approaches. In this sense, this section will present the most used datasets, frameworks and tools in the area.

7.1. Datasets

When developing a new recognition systems or improving an existing one, the datasets to test need to be chosen carefully. Dataset properties vary widely, and overfitting at model training can lead to illogical results. In the field of HBA following video datasets stand out:

- **HOHA – Hollywood human actions** (Laptev, Marszalek, Schmid, & Rozenfeld, 2008): this dataset contains video sequences from 32 movies with annotations of eight types of *actions*: *AnswerPhone*, *GetOutCar*, *HandShake*, *HugPerson*, *Kiss*, *SitDown*, *SitUp* and *StandUp*. Training and testing sets are provided, as well as an automatically labelled training set with approximately 60% correct labels. A second version is available with about 1200 min of video and four new actions in addition to the existing ones: *DriveCar*, *Eat*, *Fight* and *Run*. As video clips are taken from movies, persons in the images are focused mainly and background changes are frequent. Therefore, this dataset is very useful and challenging. Nevertheless, it should not be forgotten that this type of images is difficult to obtain with regular surveillance cameras.

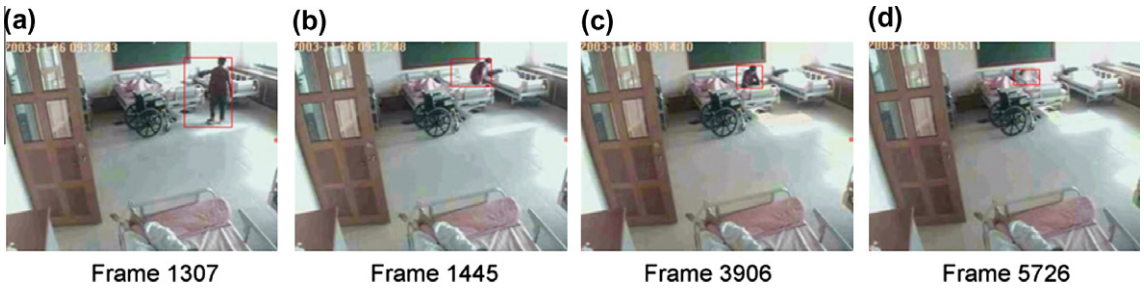

Fig. 9. Action and activity recognition from Chung and Liu (2008) which allows further behaviour recognition. Different *actions* ((a) *walking to bed*; (b) *sitting on the bed*; (c) *resting on bed*; (d) *lying on bed*) imply an *activity*: *go to sleep*, which is detected by time limitation (reprinted).

- **KTH human motion dataset** (Schuldt, Laptev, & Caputo, 2004): this *action* database contains six types of human actions performed by 25 subjects at four different scenarios. *Walking*, *jogging*, *running*, *boxing*, *hand waving* or *hand clapping* are performed in over 2000 sequences. Backgrounds are homogeneous and free of clutter. Video files are classified by actions, so that unwanted actions can be excluded easily. In contrast to the HOHA dataset, background segmentation is much easier with this type of images; and annotated actions can be placed at the same semantic abstraction level.
- **Weizmann human action dataset** (Gorelick et al., 2007): Gorelick et al. used static front-side cameras to record single human motion from 10 subjects in different environments. About 340 MB of video sequences are available; performed *actions* include *walking*, *running*, *bending*, *hand waving* and different types of *jumping*. The corresponding background sequences, with no subjects, and the subtraction masks—either with post-aligning or without it—are available too. The system is based on space–time features and is able to recognise complex actions like ballet movements.
- **INRIA Xmas motion acquisition sequences** (Weinland, Ronfard, & Boyer, 2006): This dataset includes 390×291 pixels video images recorded from five different angles. Eleven actors performed 13 *actions*: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *point*, *pick up*, *throw over head* and *throw from bottom up*. These actions were performed three times each, in an arbitrary chosen angle in relation to the view-point. Backgrounds and illumination settings are static and free of clutter.
- **TUM kitchen dataset** (Tenorth, Bandouch, & Beetz, 2009): This dataset targets ADLs at a kitchen scenario at a low *action* level. *Table setting*is performed by several subjects in different ways; some transport items one by one; and other behave natural, grasping several objects at once. Video images have a resolution of 384×288 pixels at 25 fps; and motion capture data, extracted with a marker-less full-body tracker, is provided. Furthermore, RFID tag readings from fixed readers at the placemat, the napkin, the plate and the cup; and sensor data from magnetic sensors at doors and drawers are available. Each frame has been labelled manually and separately for the left hand, the right hand and the trunk of the person. Among others, actions like *carrying an object*, *standing still*, *reaching*, *walking*, *taking something*, or *closing a door* are labelled.
- **MuHAVI dataset** (Singh, Velastin, & Ragheb, 2010): By targeting silhouette-based human action recognition methods, this dataset includes video data obtained from multiple cameras. Images are taken with night street light illumination at a constant but uneven background. At each corner and each side of a rectangular platform a *Schwan* CCTV camera is installed. These cameras captured, according to our taxonomy, 16 different *actions* (*WalkTurnBack*, *RunStop*, *Punch Kick*, *hotGunCollapse*, *PullHeavyObject*, *PickupThrowObject*, *WalkFall*, *LookInCar*, *CrawlOnKnees*, *WaveArms*, *JumpOverFence*, *DrunkWalk*, *ClimbLadder*, *SmashObject*, *JumpOverGap*) and one *activity* (*DrawGraffiti*) performed by 7 actors, three times each. Each frame has a 720×576 pixels resolution and is taken at 25 fps. Nevertheless, silhouettes are annotated only at a small sub-set of the available video data.
- **UCF sport action datasets** (Rodriguez, Ahmed, & Shah, 2008): Among other datasets available at UCF, this dataset stands out as it contains nearly 200 video sequences at a resolution of 720×480 pixels. Images are intentionally taken from real scenarios (usually from broadcast television channels), as on purpose recorded performances from actors lead to unrealistic and laboratory-conditioned training data. On the contrary, images taken from sport broadcasting; or from *Youtube*, as happens at the UCF50 dataset; present large variations in camera motion, object appearance and scale, viewpoint, clutter and illumination settings; and are therefore very challenging. Considering our taxonomy of HBA levels, this dataset does not only include *actions* (*walking*, *swinging*, *running*, *diving*, *golf swinging*, *kicking*, *lifting*), but also activities (*horseback riding*, *skating*).
- **CAVIAR test scenarios**: The CAVIAR project (CAVIAR Project, 2004) also published its database. Its images are taken in two different scenarios: an entrance lobby and a shopping center. *Activities* of real scenarios are recorded (*walking alone*, *meeting other people*, *window shopping*, *entering and exiting shops*, *fighting*, *passing out* and *leaving a package in a public place*) at a resolution of 384×288 pixels. Ground-truth data is provided in XML format at frame level. Video sequences, taken from wide angle cameras installed as surveillance cameras at the ceiling corners, include several persons, as well as crowd movements.
- **CMU-MMAC database** (De la Torre Frade et al., 2008): The multi-modal activity database from the Carnegie Mellon University targets cooking and food preparation *activities*. Not only video data has been taken, but also audio and other sensor data (motion, accelerometers and gyroscopes). Five subjects were recorded in a kitchen while preparing five different recipes: *brownies*, *pizza*, *sandwich*, *salad* and *scrambled eggs*. Video images were taken from three high spatial resolution cameras (1024×768) at low temporal resolution (30 fps) and three low spatial cameras (640×480), two at high temporal resolution (60 fps), and a wearable one at low temporal resolution (12 fps). Audio data was recorded with five balanced microphones and a wearable watch. Motion was captured with 12 infrared cameras of 4 MP at 120 fps. Five 3-axis accelerometers and gyroscopes contributed to the rest of the data. The computers used to record the sensor data were synchronised using the Network Time Protocol (NTP).
- **PlaceLab datasets** (Intille et al., 2006): The PlaceLab live-in laboratory provides a full home-like environment for data gathering for ubiquitous technologies and home settings studies. Two datasets are available; whereas PLIA1 is a legacy dataset,

PLIA2 improves data sharing and visualisation by employing new data formats. This second dataset is also compatible with their visualisation and annotation tool called Handlense. PLIA2 includes 4 h of video data (infrared and RGB), in which one subject performs common household *activities* (*preparing a recipe*, *doing a load of dishes*, *cleaning the kitchen*, *doing laundry*, *making the bed*, and *light cleaning around the apartment*). Besides video data, while performing the activities, accelerometer data is recorded by so called MITes, which are attached to objects of interest (i.e. objects which are related to human activities) as remote controls, chairs, etc. Videos are annotated not only with the type of activity, but also with body posture, location and social context.

Table 2 summarises the details of the reviewed datasets.

7.2. Frameworks and tools

This section will detail recently appeared frameworks and tools in the field of HBA and AAL in smart homes. As these fields present a lack of standards and interoperability, first steps in multipurpose design of tools; as languages, meta-models and frameworks; have been taken in the last few years:

- **Home markup language** (Nugent et al., 2007): HomeML is an XML based schema for representation of information within smart homes. As data taken at a smart home scenario belongs to heterogeneous nature, and is captured by different type of sensors; this language offers an open standard for the exchange of data in a system-, application- and format-independent way. Their ultimate goal is to support the exchange of data and to build an open data repository. HomeML supports a data structure which is designed upon the most used standards in integration of home services and devices: OSGi and KNX. This data structure is designed as a series of hierarchical data trees which enables a classified storage of the descriptions of the smart home environment (rooms, floors, inhabitants), and its devices and related events.
- **ViPER – The Video Performance Evaluation Resource** (Language & Media Processing Laboratory, University of Maryland, 2003): ViPER is a framework which targets semantic video analysis and includes several tools which make system evaluation easier. In this sense, the framework includes a *Ground Truth Authoring Tool* which incorporates a GUI to edit ground truth data and check generated metadata frame by frame. Once this step is done, performance of our recognition algorithm can be evaluated with batch-processes in a UNIX environment. In addition, a run-time application loader for JavaBeans and a Java MPEG-1 decoder with frame indexing are provided. As video metadata is stored in XML format and follows a specially designed structure, an API is provided in the form of a set of Java interfaces to access metadata programmatically; as well as a browser which visualises ground truth data and analyses results in several representation forms.
- **Hong et al.'s activity meta-model:** In Hong et al. (2009), Hong et al. present a new meta-model for activity recognition in smart homes. A diagram, which is similar to an Entity–Relation, is used to build evidential networks which express the interaction between recognised activities and objects. This way, relationships between activities and objects, as well as generalisation at activity level and compulsory or optional interaction with objects can be captured. Sensors' associations to objects and vice versa can be captured too. For instance, in Fig. 10, the activity of *making a cold drink* is associated with the composite object *cup-juice*. The objects *cup* and *juice* are compulsory to their combination, i.e. the composite object. Whereas the object cup is associated directly with the sensor called *scup*, the object *juice* is derived from the object *fridge* and this one is the object which is associated to the sensor of the fridge.
- **BehaviourScope Framework** (Bamis, Lymberopoulos, Teixeira, & Savvides, 2010): The Embedded Networks and Applications Lab at the University of Yale developed a scalable framework for detailed behaviour interpretation of the elderly. Its aim is to process, communicate and present heterogeneous sensor data in an automated form, in order to infer high-level semantic data, which can be further processed at applications and services (generation of alarms, reports, triggers and answers to queries is considered). Sensors like passive infrared, door/windows opening and cameras are supported, whereas new types of sensors can be added by developing the appropriate driver for the gateway. Cameras are not used for video streaming, but for motion detection and tracking based on a motion histogram; their aim is to include this processing in a new camera chip, that would avoid providing any image information to the rest of the system. The framework also includes a Web portal for visualisation and customisation, and a mobile phone application to provide personal safety services.
- **OpenAAL** (Wolf et al., 2010): The FZI Research Center for IT, the Friedrich–Schiller University of Jena and the CAS Software AG released this open source middleware for AAL last year. OpenAAL has been developed since 2007, as it started as the technical development of the SOPRANO Integrated Project (Sixth Framework Programme of IST) (Wolf, Schmidt, & Klein, 2008). On top of the OSGi service-oriented framework, OpenAAL provides generic platform services based on three main components: (1) the Context Manager, where ambient data and information from sensors and user inputs are collected and stored supporting context reasoning at multiple levels of abstraction (from sensor and actuator states to environment characteristics); (2)

Table 2
Comparison of dataset features.

Dataset	DoS	'Actions'	Multi-view	Maximum resolution	Background	Silhouettes	Out-/indoor
HOHA	Actions	8/12	No	240 lines	Complex	No	*both*
KTH	Actions	6	No	160 × 120	Simple	No	*both*
Weizmann	Actions	10	No	180 × 144	Simple	Yes	outdoor
INRIA-XMAS	Actions	13	Yes	390 × 291	Simple	Yes	indoor
TUM Kitchen	Actions	10[a]	Yes	780 × 582	Simple	No	indoor
MuHAVI	*Both*	17	Yes	720 × 576	Complex	Yes[b]	indoor
UCF Sports	*Both*	9	No	720 × 480	Complex	No	*Both*
CAVIAR	Activities	6	Yes	384 × 288	Complex	No	Indoor
CMU-MMAC	Activities	5	Yes	1024 × 768	Simple	No	Indoor
PlaceLab (PLIA2)	Activities	6	Yes	320 × 240	Simple	No	Indoor

[a] Approximately 10 annotated sub-actions of 1 activity: setting the table.
[b] They are provided for the Manually-Annotated Subset (MAS).

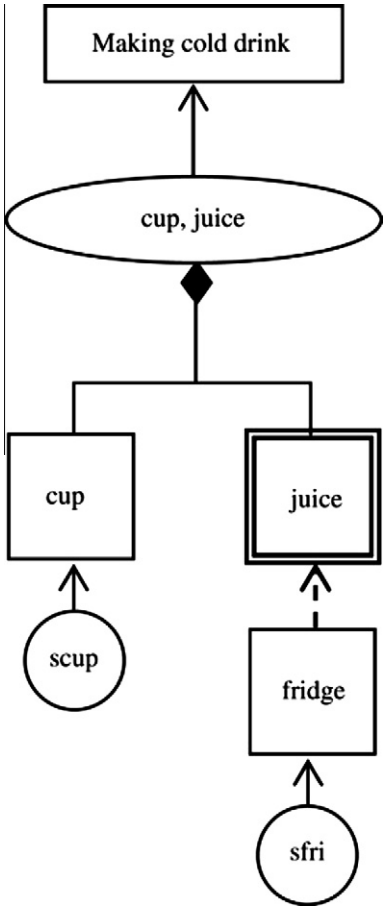

Fig. 10. Evidential network representation of making a cold drink. Reprinted from Hong et al. (2009), ©Elsevier, 2008.

The Procedural Manager which is in charge of handling installation-independent workflows which are able to react to situations of interest. These workflows are defined in BPEL with context-aware extensions in order to be able to communicate with the Context Manager; and (3) The Composer selects the available services in the concrete installation to achieve the abstract service goals; these are described in the installation-independent workflows. This way, abstract services can be concreted with the appropriate combination of services in order to adapt to the user's needs in each situation.

The middleware is available online with LGPL licence and documentation is provided. The developed code is written in Java and uses the open source implementation of the OSGi R4 core framework specification Equinox.

8. Conclusions

This paper has covered the different levels of HBA following an abstraction, degree of semantics and time oriented classification. Going through recent examples of research works; most used and promising feature types, recognition methods and system design methodologies have been detailed. From pose, gaze and motion estimation to behaviour recognition, and following an initially defined classification, we have analysed vision and multi-modal-based approaches.

Clearly, it can be seen that at the motion, pose and gaze estimation level, several methods achieve robust and high success rates. In conjunction with the action level, these show the most advanced and successful results. Nevertheless at higher levels, especially at behaviour, there is still a long way to go to achieve off-the-shelf products. Still, huge advances have been made in the last ten years. But the challenge to design and develop stable and general systems still persists, as most systems only solve specific problems in very particular environments.

Especially at the field of AAL, advances in this works are very valuable as personal autonomy and quality of life for elderly and cognitively impaired people can be improved enormously by these systems, and at the same time care costs can be reduced significantly.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation under project "Sistema de visión para la monitorización de la actividad de la vida diaria en el hogar" (TIN2010-20510-C04-02). Alexandros Andre Chaaraoui acknowledges financial support by the Conselleria d'Educació, Formació i Ocupació of the Generalitat Valenciana (fellowship ACIF/2011/160). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Agarwal, A., & Triggs, B. (2004). 3d human pose from silhouettes by relevance vector regression. In *Proceedings of the 2004 IEEE Computer Society conference on computer vision and pattern recognition, 2004. CVPR 2004* (pp. II-882–II-888, Vol. 2).

Altun, K., & Barshan, B. (2010). Human activity recognition using inertial/magnetic sensor units. In A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human behavior understanding. Lecture notes in computer science* (Vol. 6219, pp. 38–51). Berlin/Heidelberg: Springer.

Aminian, K., Robert, P., Buchser, E., Rutschmann, B., Hayoz, D., & Depairon, M. (1999). Physical activity monitoring based on accelerometry: Validation and comparison with video observation. *Medical and Biological Engineering and Computing, 37*, 304–308.

Anderson, D., Luke, R., Keller, J., Skubic, M., Rantz, M., & Aud, M. (2009). Modeling human activity from voxel person using fuzzy logic. *IEEE Transactions on Fuzzy Systems, 17*, 39–49.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 1014–1021).

Bamis, A., Lymberopoulos, D., Teixeira, T., & Savvides, A. (2010). The behaviorscope framework for enabling ambient assisted living. *Personal and Ubiquitous Computing, 14*, 473–487.

Bandouch, J., Engstler, F., & Beetz, M. (2008). Accurate human motion capture using an ergonomics-based anthropometric human model. In F. Perales & R. Fisher (Eds.), *Articulated motion and deformable objects. Lecture notes in computer science* (Vol. 5098, pp. 248–258). Berlin/Heidelberg: Springer.

Bao, L., & Intille, S. (2004). Activity recognition from user-annotated acceleration data. In A. Ferscha & F. Mattern (Eds.), *Pervasive computing. Lecture notes in computer science* (Vol. 3001, pp. 1–17). Berlin/Heidelberg: Springer.

Beetz, M., Bandouch, J., Jain, D., & Tenorth, M. (2010). Towards automated models of activities of daily life. *Technology and Disability, 4*, 1–11.

Belongie, S., Malik, J., & Puzicha, J. (2000). Shape context: A new descriptor for shape matching and object recognition. In *Proceedings of the neural information processing systems 2000* (pp. 831–837).

Ben-Arie, J., Wang, Z., & Pandit, P. (2002). Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 1091–1104.

Bobick, A., & Davis, J. (2001). The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, 23*, 257–267.

Boulay, B., Bremond, F., & Thonnat, M. (2006). Applying 3d human model in a posture recognition system. *Pattern Recognition Letters, 27*, 1788–1796.

Bourdev, L., Maji, S., Brox, T., & Malik, J. (2010). Detecting people using mutually consistent poselet activations. *Proceedings of the 11th European conference on Computer vision: Part VI ECCV'10* (pp. 168–181). Berlin, Heidelberg: Springer-Verlag.

Brand, M. (1996). Coupled hidden markov models for modeling interacting processes. *Neural Computation, 405*, 1–28.

Brdiczka, O., Crowley, J., & Reignier, P. (2009). Learning situation models in a smart home. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39*, 56–63.

Canton-Ferrer, C., Segura, C., Pardas, M., Casas, J., & Hernando, J. (2008). Multimodal real-time focus of attention estimation in smartrooms. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08* (pp. 1–8).

Cardinaux, F., Brownsell, S., Hawley, M., & Bradley, D. (2008). Modelling of behavioural patterns for abnormality detection in the context of lifestyle

reassurance. In J. Ruiz-Shulcloper & W. Kropatsch (Eds.), *Progress in pattern recognition, image analysis and applications. Lecture notes in computer science* (Vol. 5197, pp. 243–251). Berlin/Heidelberg: Springer.

Carranza, J., Theobalt, C., Magnor, M. A., & Seidel, H.-P. (2003). Free-viewpoint video of human actors. In *ACM SIGGRAPH 2003 papers SIGGRAPH '03* (pp. 569–577). New York, NY, USA: ACM.

CAVIAR Project (2004). Caviar test case scenarios. <http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1>. Accessed 18.01.12.

Cheng, H., Yang, C., Han, F., & Sawhney, H. (2008). Ho2: A new feature for multi-agent event detection and recognition. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08* (pp. 1–8).

Cherla, S., Kulkarni, K., Kale, A., & Ramasubramanian, V. (2008). Towards fast, view-invariant human action recognition. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08* (pp. 1–8).

Cheung, K., Baker, S., & Kanade, T. (2003). Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society conference on computer vision and pattern recognition, 2003* (Vol. 1, pp. I-77–I-84).

Chung, P.-C., & Liu, C.-D. (2008). A daily behavior enabled hidden markov model for human behavior understanding. *Pattern Recognition, 41*, 1572–1580.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05). CVPR '05* (Vol. 1, pp. 886–893). Washington, DC, USA: IEEE Computer Society.

De la Torre Frade, F., Hodgins, J. K., Bargteil, A. W., Martin Artal, X., Macey, J. C., Collado I Castells, A., et al. (2008). Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. Technical Report CMU-RI-TR-08-22 Robotics Institute Pittsburgh, PA.

Dempster, A. P. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological), 30*, 205–247.

Doshi, A., & Trivedi, M. (2010a). Attention estimation by simultaneous observation of viewer and view. In *2010 IEEE Computer Society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 21–27). IEEE.

Doshi, A., & Trivedi, M. (2010b). Examining the impact of driving style on the predictability and responsiveness of the driver: Real-world and simulator analysis. In *2010 IEEE intelligent vehicles symposium (IV)* (pp. 232–237). IEEE.

Duong, T., Bui, H., Phung, D., & Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE Computer Society conference on computer vision and pattern recognition, 2005. CVPR 2005* (Vol. 1, pp. 838–845).

Eichner, M., & Ferrari, V. (2009). Better appearance models for pictorial structures. In *Proceedings of the British machine vision conference* (pp. 1–11).

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 1627–1645.

Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8).

Gavrila, D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding, 73*, 82–98.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space–time shapes. *Transactions on Pattern Analysis and Machine Intelligence, 29*, 2247–2253.

Hara, K., Omori, T., & Ueno, R. (2002). Detection of unusual human behavior in intelligent house. In *Proceedings of the 2002 12th IEEE workshop on neural networks for signal processing, 2002* (pp. 697–706).

Hayes, T., Pavel, M., & Kaye, J. (2008). An approach for deriving continuous health assessment indicators from in-home sensor data. In *Technology and aging: Selected papers from the 2007 international conference on technology and aging* (Ios Press Vol. 21 pp. 130–137).

Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., & Devlin, S. (2009). Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing, 5*, 236–252.

Hongeng, S., & Nevatia, R. (2001). Multi-agent event recognition. *IEEE international conference on computer vision. 2001 Eighth ICCV 2001* (Vol. 2, pp. 84–91). IEEE.

Howe, N. (2004). Silhouette lookup for automatic pose tracking. In *Conference on computer vision and pattern recognition workshop, 2004. CVPRW '04* (pp. 15–22).

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 34*, 334–352.

Intille, S. S., Larson, K., Tapia, E. M., Beaudin, J., Kaushik, P., Nawyn, J., et al. (2006). Using a live-in laboratory for ubiquitous computing research. In *Proceedings of PERVASIVE06* (pp. 349–365).

Jaimes, A., & Sebe, N. (2007). Multimodal human computer interaction: A survey. *Computer Vision and Image Understanding, 108*, 116–134.

Jain, G., Cook, D., & Jakkula, V. (2006). Monitoring health by detecting drifts and outliers for a smart environment inhabitant. In *Proceedings of the international conference on smart homes and health telematics* (pp. 1–8).

Ji, X., Liu, H., Li, Y., & Brown, D. (2008). Visual-based view-invariant human motion analysis: A review. In I. Lovrek, R. Howlett, & L. Jain (Eds.), *Knowledge-based intelligent information and engineering systems. Lecture notes in computer science* (Vol. 5177, pp. 741–748). Berlin/Heidelberg: Springer.

Juan, L., & Gwon, O. (2009). A Comparison of SIFT, PCA-SIFT and SURF. *International Journal of Image Processing (IJIP), 3*, 143–152.

Kadir, T., & Brady, M. (2003). Scale saliency: A novel approach to salient feature and scale selection. In *International conference on visual information engineering, 2003. VIE 2003* (pp. 25–28). IET.

Karaman, S., Benois-Pineau, J., Mégret, R., Dovgalecs, V., Dartigues, J., & Gäestel, Y. (2010). Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *2010 International conference on pattern recognition* (pp. 4113–4116). IEEE.

Kwapisz, J., Weiss, G., & Moore, S. (2011). Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter, 12*, 74–82.

Language and Media Processing Laboratory, University of Maryland (2003). Viper: The video performance evaluation resource. <http://viper-toolkit.sourceforge.net>. Last accessed 18.01.12.

Laptev, I. (2005). On space–time interest points. *International Journal of Computer Vision, 64*, 107–123.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8).

Launila, A., & Sullivan, J. (2010). Contextual features for head pose estimation in football games. In *2010 20th International conference on pattern recognition* (ICPR), (pp. 340–343).

Lester, J., Choudhury, T., & Borriello, G. (2006). A practical approach to recognizing physical activities. In *Proceedings of the pervasive* (pp. 1–16).

Li, W., Zhang, Z., & Liu, Z. (2008). Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology, 18*, 1499–1510.

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 9–14).

Liu, C.-D., Chung, Y.-N., & Chung, P.-C. (2010). An interaction-embedded hmm framework for human behavior understanding: With nursing environments as examples. *IEEE Transactions on Information Technology in Biomedicine, 14*, 1236–1246.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91–110.

Lucas, B. D. (1981). An iterative image registration technique with an application to stereo vision. *Imaging, 130*, 121–129.

Lv, F., & Nevatia, R. (2007). Single view human action recognition using key pose matching and viterbi path searching. In *IEEE conference on computer vision and pattern recognition, 2007. CVPR '07* (pp. 1–8).

Mahmoud, S., Akhlaghinia, M., Lotfi, A., & Langensiepen, C. (2011). Trend modelling of elderly lifestyle within an occupancy simulator. In *2011 UkSim 13th international conference on computer modelling and simulation (UKSim)* (pp. 156–161).

Marin-Jimenez, M., Zisserman, A., & Ferrari, V. (2011). "Here's looking at you, kid." Detecting people looking at each other in videos. In *Proceedings of the British machine vision conference*.

Maurer, U., Smailagic, A., Siewiorek, D., & Deisher, M. (2006). Activity recognition and monitoring using multiple sensors on different body positions. In *International workshop on wearable and implantable body sensor networks, 2006. BSN 2006* (pp. 113–116).

Mcivor, A. M. (2000). Background subtraction techniques. In *Proceedings of the image and vision computing, Auckland, New Zealand, 2000*.

Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th international conference on computer vision* (pp. 104–111).

van der Meulen, P., & Seidl, A. (2007). Ramsis the leading cad tool for ergonomic analysis of vehicles. In V. Duffy (Ed.), *Digital human modeling. Lecture notes in computer science* (Vol. 4561, pp. 1008–1017). Berlin/Heidelberg: Springer.

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

Mihailidis, A., Carmichael, B., & Boger, J. (2004). The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Transactions on Information Technology in Biomedicine, 8*, 238–247.

Mihailidis, A., Fernie, G. R., & Cleghorn, W. L. (2000). The development of a computerized cueing device to help people with dementia to be more independent. *Technology and Disability, 13*, 23–40.

Mikić, I., Trivedi, M., Hunter, E., & Cosman, P. (2003). Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision, 53*, 199–223.

Moeslund, T. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding, 81*, 231–268.

Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding, 104*, 90–126.

Monekosso, D. N., & Remagnino, P. (2010). Behavior analysis for assisted living. *IEEE Transactions on Automation Science and Engineering, 7*, 879–886.

Nait-Charif, H., & McKenna, S. (2004). Activity summarisation and fall detection in a supportive home environment. In *Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004* (Vol. 4, pp. 323–326).

Najafi, B., Aminian, K., Paraschiv-Ionescu, A., Loew, F., Büla, C. J., & Robert, P. (2003). Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly. *IEEE Transactions on Bio-Medical Engineering, 50*, 711–723.

Nguyen, N., Phung, D., Venkatesh, S., & Bui, H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. In *IEEE Computer Society conference on computer vision and pattern recognition, 2005. CVPR 2005* (Vol. 2, pp. 955–960).

Nicolini, C., Lepri, B., Teso, S., & Passerini, A. (2010). From on-going to complete activity recognition exploiting related activities. In A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human behavior understanding. Lecture notes in computer science* (Vol. 6219, pp. 26–37). Berlin/Heidelberg: Springer.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision, 79*, 299–318.

Nugent, C. D., Finlay, D. D., Davies, R. J., Wang, H. Y., Zheng, H., Hallberg, J., et al. (2007). homeml: An open standard for the exchange of data within smart environments. In *Proceedings of the 5th international conference on smart homes and health telematics ICOST'07* (pp. 121–129). Berlin, Heidelberg: Springer-Verlag.

Oikonomopoulos, A., Pantic, M., & Patras, I. (2008). B-spline polynomial descriptors for human activity recognition. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08* (pp. 1–6).

Oikonomopoulos, A., Patras, I., & Pantic, M. (2005). Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36*, 710–719.

Oikonomopoulos, A., Patras, I., & Pantic, M. (2009). An implicit spatiotemporal shape model for human activity localization and recognition. In *2009 IEEE Computer Society conference on computer vision and pattern recognition workshops* (pp. 27–33).

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 831–843.

Ozturk, O., Yamasaki, T., & Aizawa, K. (2009). Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *2009 IEEE 12th international conference on computer vision workshops (ICCV Workshops)* (pp. 1020–1027).

Park, K., Lin, Y., Metsis, V., Le, Z., & Makedon, F. (2010). Abnormal human behavioral pattern detection in assisted living environments. In *Proceedings of the 3rd international conference on pervasive technologies related to assistive environments PETRA '10* (pp. 9:1–9:8). New York, NY, USA: ACM.

Park, S., & Kautz, H. (2008). Privacy-preserving recognition of activities in daily living from multi-view silhouettes and rfid-based training. In *AAAI fall 2008 symposium on AI in Eldercare: New solutions to old problems*.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding, 108*, 4–18.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*, 976–990.

Porle, R. R., Chekima, A., Wong, F., & Sainarayanan, G. (2009). Performance of histogram-based skin colour segmentation for arms detection in human motion analysis application. *International Journal of Electronics, Communications and Computer Engineering, 1*, 403–408.

Quintas, J., Khoshhal, K., Aliakbarpour, H., Hofmann, M., & Dias, J. (2011). Using concurrent hidden markov models to analyse human behaviours in a smart home environment. In *12th International workshop on image analysis for multimedia interactive services*.

Reale, M., Hung, T., & Yin, L. (2010). Pointing with the eyes: Gaze estimation using a static/active camera system and 3d iris disk model. In *2010 IEEE international conference on multimedia and expo (ICME)* (pp. 280–285). IEEE.

Reale, M., Hung, T., & Yin, L. (2010). Viewing direction estimation based on 3d eyeball construction for hri. In *2010 IEEE Computer Society conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 24–31). IEEE.

Ren, X., & Philipose, M. (2009). Egocentric recognition of handled objects: Benchmark and analysis. In *IEEE Computer Society conference on computer vision and pattern recognition workshops* (pp. 1–8). IEEE.

Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding, 104*, 232–248.

Robertson, N., Reid, I., & Brady, M. (2006). Behaviour recognition and explanation for video surveillance. In The institution of engineering and technology conference on crime and security, 2006 (pp. 458–463).

Rodriguez, M., Ahmed, J., & Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8).

Roetenberg, D., Slycke, P. J., & Veltink, P. H. (2007). Ambulatory position and orientation tracking fusing magnetic and inertial sensing. *IEEE Transactions on Bio-Medical Engineering, 54*, 883–890.

Rosales, R., & Sclaroff, S. (2000). Learning and synthesizing human body motion and posture. In *Fourth IEEE international conference on automatic face and gesture recognition, 2000* (pp. 506 –511).

Rybok, L., Voit, M., Ekenel, H., & Stiefelhagen, R. (2010). Multi-view based estimation of human upper-body orientation. In 2010 20th International conference on pattern recognition (ICPR) (pp. 1558–1561).

Ryoo, M., & Aggarwal, J. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In 2009 IEEE 12th international conference on computer vision (pp. 1593–1600).

Sapp, B., Toshev, A., & Taskar, B. (2010). Cascaded models for articulated pose estimation. In *Proceedings of the 11th European conference on computer vision: Part II ECCV'10* (pp. 406–420). Berlin, Heidelberg: Springer-Verlag.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004* (Vol. 3, pp. 32–36).

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In *Ninth IEEE international conference on computer vision, 2003* (Vol. 2, pp. 750–757).

Shimizu, H., & Poggio, T. (2003). Direction estimation of pedestrian from images.

Shotton, J., Fitzgibbon, A. W., Cook, M., Sharp, T., Finocchio, M., Moore, R., et al. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 24th IEEE conference on computer vision and pattern recognition, CVPR 2011* (pp. 1297–1304).

Singh, S., Velastin, S., & Ragheb, H. (2010). Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 Seventh IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 48–55). IEEE.

Sminchisescu, C., Kanaujia, A., Li, Z., & Metaxas, D. (2005). Discriminative density propagation for 3d human motion estimation. In *IEEE Computer Society conference on computer vision and pattern recognition, 2005. CVPR 2005* (Vol. 1, pp. 390–397).

Spriggs, E., De La Torre, F., & Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2009. CVPR Workshops 2009* (pp. 17–24).

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Computer Society conference on computer vision and pattern recognition, 1999* (Vol. 2, pp. 246–252).

Sun, L., Klank, U., & Beetz, M. (2009). Eyewatchme3d hand and object tracking for inside out activity analysis. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2009. CVPR workshops 2009* (pp. 9–16). IEEE.

Sundaram, S., & Cuevas, W. (2009). High level activity recognition using low resolution wearable vision. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2009. CVPR workshops 2009* (pp. 25–32).

Sutton, C., & McCallum, A. (2007). An introduction to conditional random fields for relational learning. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning*. MIT Press.

Tao, Y., Hu, H., & Zhou, H. (2007). Integration of vision and inertial sensors for 3d arm motion tracking in home-based rehabilitation. *The International Journal of Robotics Research, 26*, 607–624.

Tapia, E., Intille, S., & Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors pervasive computing. In A. Ferscha & F. Mattern (Eds.), *Pervasive computing. Lecture notes in computer science* (Vol. 3001, pp. 158–175). Berlin, Heidelberg: Springer.

Tenorth, M., Bandouch, J., & Beetz, M. (2009). The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *IEEE 12th international conference on computer vision workshops (ICCV Workshops), 2009* (pp. 1089 –1096).

Toyama, K., Krumm, J., Brumitt, B., & Meyers, B. (1999). Wallflower: Principles and practice of background maintenance. In *The Proceedings of the seventh IEEE international conference on computer vision, 1999* (Vol. 1, pp. 255–261).

Turaga, P., Chellappa, R., Subrahmanian, V., & Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology, 18*, 1473–1488.

Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision, 3*, 177–280.

Virone, G., & Sixsmith, A. (2008). Monitoring activity patterns and trends of older adults. In *30th Annual international conference of the IEEE engineering in medicine and biology society, 2008. EMBS 2008* (pp. 2071–2074).

Wang, L. (2003). Recent developments in human motion analysis. *Pattern Recognition, 36*, 585–601.

Weinland, D., Boyer, E., & Ronfard, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *IEEE 11th international conference on computer vision, 2007. ICCV 2007* (pp. 1–7).

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding, 104*, 249–257.

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Tenth IEEE international conference on computer vision, 2005. ICCV 2005* (Vol. 2, pp. 1800–1807).

Wolf, P., Schmidt, A., & Klein, M. (2008). Soprano – An extensible, open aal platform for elderly people based on semantical contracts. In *3rd Workshop on artificial intelligence techniques for ambient intelligence (AITAmI08), 18th European conference on artificial intelligence (ECAI 08), Patras, Greece*.

Wolf, P., Schmidt, A., Otte, J. P., Klein, M., Rollwage, S., Knig-Ries, B., Dettborn, T., et al. (2010). openaal – The open source middleware for ambient-assisted living (aal). In *AALIANCE conference, Malaga, Spain, March 11–12, 2010*.

Wood, A., Stankovic, J., Virone, G., Selavo, L., He, Z., Cao, Q., et al. (2008). Context-aware wireless sensor networks for assisted living and residential monitoring. *IEEE Network, 22*, 26–33.

Wu, B., & Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In *Tenth IEEE international conference on computer vision, 2005. ICCV 2005* (Vol. 1, pp. 90–97).

Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., & Rehg, J. (2007). A scalable approach to activity recognition based on object use. In *IEEE 11th international conference on computer vision, 2007. ICCV 2007* (pp. 1–8). IEEE.

Xu, C., & Prince, J. (1998). Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing, 7*, 359–369.

Yamamoto, M., Mitomi, H., Fujiwara, F., & Sato, T. (2006). Bayesian classification of task-oriented actions based on stochastic context-free grammar. In *7th international conference on automatic face and gesture recognition, 2006. FGR 2006* (pp. 317–322).

Yang, A., Iyengar, S., Sastry, S., Bajcsy, R., Kuryloski, P., & Jafari, R. (2008). Distributed segmentation and classification of human actions using a wearable motion sensor network. In *IEEE Computer Society conference on computer vision and pattern recognition workshops, 2008. CVPRW '08* (pp. 1–8).

Zhou, Z., Chen, X., Chung, Y.-c., He, Z., Han, T. X., & Keller, J. M. (2008). Activity analysis, summarization, and visualization for indoor human activity monitoring. *IEEE Transactions on Circuits and Systems for Video Technology, 18*, 1489–1498.

Zouba, N., Boulay, B., Bremond, F., & Thonnat, M. (2008). *Cognitive vision. Chapter monitoring activities of daily living (ADLs) of elderly based on 3D key human postures*. Berlin, Heidelberg: Springer-Verlag.