Useful Research Tools for Human Behaviour Understanding in the context of Ambient Assisted Living

Pau Climent-Pérez, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta

Abstract When novice researchers in the fields of Computer Vision and Human Behaviour Understanding initiate new projects applied to Ambient-Assited Living (AAL) scenarios, a lack of specific, publicly available frameworks, tools and datasets is perceived. This work is an attempt to fill that particular gap, by presenting different field-related datasets—or benchmarks—, according to a taxonomy (which is also presented), and taking into account their availability as well as their relevance. Furthermore, it reviews and puts together a series of tools—either frameworks or pieces of software—that are at hand (although dispersed), which can ease the task. To end with the work, some conclusions are drawn about the reviewed tools, putting special emphasis in their generality and reliability.

1 Introduction

When initiating oneself in the world of Computer Vision, and more specifically in Human Behaviour Analysis—and Understanding—(HBA/HBU) by means of vision devices, things are particularly complicated; many research is published, but not accompanied with publicly available datasets or tools employed in it, which renders reproduction of results a very time-consuming task.

Therefore, the present work was conceived for those who are starting 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 work will review the most used datasets, frameworks and tools in the area.

A wide variety of datasets exists, but they can be grouped according to the semantic level of the tasks being recognised, thus achieving a four-level taxonomy that will be further explained.

Pau Climent-Pérez · Alexandros Andre Chaaraoui · Francisco Flórez-Revuelta
Department of Computing Technology, University of Alicante, Ctra. San Vicente del Raspeig, s/n, 03690 San Vicente del Raspeig, Alicante, Spain e-mail: {pcliment, alexandros, florez}@dtic.ua.es

2 Datasets

In this paper, a taxonomy is followed, which is based on others seen in the literature such as [10, 12]. Under this taxonomy, 'actions' are classified into increasing levels of semantic richness and the time involved in the analysis. Therefore, the following degrees are presented:

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[a] 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)

[a] ADLs stands for 'Activities of Daily Living'.

In the research of new methods, datasets need to be chosen carefully according to the tasks to recognise and their degree of semantics (DoS). By using the classification in Table 1, datasets can be presented according to that specific degree of semantics which allows a better understanding of the applications of each of them. In the following, the terms *motion, action, activity* and *behaviour* refer to the semantic levels just presented. According to this, and having ADL recognition and AAL as targets, the following video datasets stand out:

HOHA - Hollywood human actions [9]: This dataset contains video sequences from 32 movies with annotations of 8 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 minutes 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.

KTH human motion dataset [14]: 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 at 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 [6]: Gorelick et al. used static front-side cameras to record single human motion from 9 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 developed in [6] is based on space-time features and is able to recognise complex actions like ballet movements.

INRIA Xmas motion acquisition sequences [17]: This dataset includes 390×291 pixels video images recorded from five different angles. 11 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 [16]: 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 labeled 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 labeled.

MuHAVI dataset [15]: 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 16 different composite actions (*WalkTurnBack, RunStop, Punch, Kick, hotGunCollapse, PullHeavyObject, PickupThrowObject, WalkFall, LookInCar, CrawlOnKnees, WaveArms, JumpOverFence, DrunkWalk, ClimbLadder, SmashObject, JumpOverGap*) and one highly complex 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 on a small subset of the available video data.

UCF Sports Action Dataset [13]: Among other datasets available at UCF[1], 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

[1] `http://server.cs.ucf.edu/~vision/data.html`

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 [5]: The CAVIAR project also published its database. Its images are taken at two different scenarios: an entrance lobby and a shopping centre. Activities of real scenarios are recorded (walking alone, meeting other people, window shopping, entering and exiting shops, fighting and 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 [4]: This is a multi-modal activity database from the Carnegie Mellon University that 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-axial 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 [8]: 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[2]; whereas PLIA1 is a legacy dataset, PLIA2 improves data sharing and visualization by employing new data formats. This second dataset is also compatible with their visualization and annotation tool called *Handlense*[3]. PLIA2 includes 4 hours 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.

[2] `http://architecture.mit.edu/house_n/data/PlaceLab/PlaceLab.htm`

[3] `http://architecture.mit.edu/house_n/data/PlaceLab/HandLense.htm`

In order to compare the characteristics of these datasets and point out their main differences, the following comparison (see Table 2) evaluates the datasets with respect to the most relevant properties. These properties have been chosen having in mind possible constraints of human behaviour analysis methods.

Table 2 Comparison of dataset features

Dataset	Taxonomy level	'Actions'	Multi-view	Maximum resolution	Background type	Silhou-ettes	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	780×582 384×288	simple	No	indoor
MuHAVI	Actions & Activities	17	Yes	720×576	complex	Yes[b]	indoor
UCF Sports	Actions & Activities	9	No	720×480	complex	No	both
CAVIAR	Activities	6	Yes	384×288	complex	No	indoor
CMU-MMAC	Activities	5	Yes	1024×768 640×480	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 in the Manually-Annotated Subset (MAS).

3 Frameworks and Tools

This section briefly analyses useful frameworks and tools for the development of AAL solutions. Only recently, generalistic, interoperation-enabling approaches have been published. Some of these first steps in multipurpose design of tools; such as languages, meta-models and frameworks are presented here.

Home markup language [11]: 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 [2] and KNX [1]. This data structure is designed as a series of hierarchical data trees which enable 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[4]**:** 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 includes 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 recognition meta-model: In [7], 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 generalization 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, 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 [3]: 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 visualization and customization, and a mobile phone application to provide personal safety services.

OpenAAL [19]: 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

[4] `http://viper-toolkit.sourceforge.net/`

Programme of IST) [18]. On top of the OSGi service-oriented framework, OpenAAL provides generic platform services based on three main components:

1. 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. The Procedural Manager 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.
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 concretised with the appropriate combination of services in order to adapt to the user's needs in each situation.

The middleware is available online[5] with LGPL license 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[6].

4 Conclusions

This paper has covered the most used datasets in the field of HBA and AAL scenarios, classifying them according to a previously defined taxonomy. It has been observed that dataset properties vary widely in respect to the quality of their images, additional data and characteristics of the environment. Advantages and particular difficulties of each dataset have been pointed out in order to ease an election.

The performed analysis on the available frameworks and tools shows that few of these have been released for public use and most research projects of this kind are unfortunately not available or, sometimes, discontinued.

Even if datasets can serve as useful benchmarks, the current goal of HBA is to develop stable and general systems that achieve successful results with changing data conditions; as most systems only solve specific problems in very particular environments. Finally, the presented frameworks and tools are helpful for speeding up repetitive development stages and, more importantly, to reach a common approach among researchers in the field.

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).

[5] `http://www.openaal.org`

[6] `http://www.eclipse.org/equinox/`

References

1. KNX Association: Handbook for Home and Building Control (2009). URL http://www.konnex.org
2. OSGi Alliance (2009). URL http://www.osgi.org
3. Bamis, A., Lymberopoulos, D., Teixeira, T., Savvides, A.: The BehaviorScope framework for enabling ambient assisted living. Personal and Ubiquitous Computing **14**(6), 473–487 (2010)
4. De la Torre, F., Hodgins, J., Montano, J., Valcarcel, S., Macey, J.: Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. Tech. rep. (2009)
5. Fisher, R.: CAVIAR Test Case Scenarios (2007). URL http://http//groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/
6. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(12), 2247–2253 (2007)
7. Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., Devlin, S.: Evidential fusion of sensor data for activity recognition in smart homes. Pervasive and Mobile Computing **5**(3), 236–252 (2009)
8. Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., Rockinson, R.: Using a live-in laboratory for ubiquitous computing research. In: K. Fishkin, B. Schiele, P. Nixon, A. Quigley (eds.) Pervasive Computing, vol. 3968, pp. 349–365. Springer (2006)
9. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conf. on, pp. 1–8. IEEE (2008)
10. Moeslund, T., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding **104**(2-3), 90–126 (2006)
11. Nugent, C., Finlay, D., Davies, R., Wang, H., Zheng, H., Hallberg, J., Synnes, K., Mulvenna, M.: homeML – an open standard for the exchange of data within smart environments. In: T. Okadome, T. Yamazaki, M. Makhtari (eds.) Pervasive Computing for Quality of Life Enhancement, vol. 4541, pp. 121–129. Springer (2007)
12. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010)
13. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conf. on, pp. 1–8. IEEE (2008)
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Pattern Recognition, 2004. ICPR 2004. Procs. of the 17th Int. Conf. on, vol. 3, pp. 32–36. IEEE (2004)
15. Singh, S., Velastin, S., Ragheb, H.: MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods. In: Advanced Video and Signal Based Surveillance (AVSS), 2010 7th IEEE Int. Conf. on, pp. 48–55. IEEE (2010)
16. Tenorth, M., Bandouch, J., Beetz, M.: The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th Int. Conf. on, pp. 1089–1096. IEEE (2009)
17. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104**(2-3), 249–257 (2006)
18. Wolf, P., Schmidt, A., Klein, M.: SOPRANO – An extensible, open AAL platform for elderly people based on semantical contracts. In: 3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'08), 18th European Conference on Artificial Intelligence (ECAI 08). Patras, Greece (2008)
19. Wolf, P., Schmidt, A., Otte, J., Klein, M., Rollwage, S., König-Ries, B., Dettborn, T., Gabdulkhakova, A.: OpenAAL – the open source middleware for ambient-assisted living (AAL). In: AALIANCE Conf., Malaga, Spain, March 11-12 (2010)