D2.3

Training Data, Collection & Annotation

Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support

Dem@Care - FP7-288199

Deliverable Information

Project Ref. No.		FP7-288199
Project Acronym		Dem@Care
Project Full Title		Dementia Ambient Care: Multi-Sensing Monitoring for Intelligence Remote Management and Decision Support.
Dissemination level:		Public
Contractual date of delivery:		30-06-2012
Actual date of delivery:		10-07-2012
Deliverable No.		D2.3
Deliverable Title		Training Data Collection & Annotation
Type:		Report
Approval Status:		Final
Version:		v16
Number of pages:		91
WP:		WP2 Requirements and Impact
Task:		Training Data Collection & Annotation
WP/Task responsible:		WP2/CHUN
Authors (Partner)		Joumier V. (CHUN), Fernando Crispim Junior C. (INRIA), Boulay B. (INRIA), Zaidenberg S. (INRIA), Bilinski P. (INRIA), Fosty B. (INRIA), Mégret R. (UB1), Schuijers E.(Philips), Newman E. (DCU), Alexander Sorin (IBM), Aharon Satt (IBM), Athina Kokonozi (CERTH).
Responsible Author	**Name**	Joumier Veronique
	Email	joumier.v@chu-nice.fr
EC Project Officer		Gisele Roesems-Kerremans. Griet van Caenegem.
Abstract (for dissemination)		This deliverable presents processes and tools for training sensor data collection and manual annotation, supporting the analysis and interpretation tasks of WP3 and WP4.

Version Log

Version	Date	Change	Author
V01	10/04/2012	First version	V. Joumier (CHUN)
V02	17/04/2012	Add ANVIL description	S. Zaidenberg (INRIA)
V03	11/05/2012	Add Viper description	B. Boulay (INRIA)
V04	23/05/2012	Add ViSEvAl description Modify Viper illustration	B. Boulay (INRIA)
V05	12/06/2012	Add SenseCam/Accelerometers/Physiological sensors description; Add training dataset description.	E. Newman (DCU), E.Schuijers (Philips)
V06	13/06/2012	Add 3D video camera Kinect sensor description	B. Fosty (INRIA)
V07	13/06/2012	Add 2D video camera sensor description/ Training data Set	C. Fernando-Crispim Junior (INRIA)
V08	21/06/2012	Add Wearable Camera Specification; Add iVi interface for annotation and consultation; Add video data corpora information; Add needs for wearable video	R. Megret (UB1)
V09	21/06/2012	Add Data Collection / Benchmarking data Set for 2D video camera	P. Bilinski (INRIA)
V10	22/06/2012	Add BBox annotation tool	J. Benois-Pineau (UB1)
V11	25/06/2012	Add Objectives and Challenges (for parts: Annotation tools and Visualisation tools)	B.Boulay (INRIA)
V12	26/06/2012	Add illustrations for Wearable Camera Specification	R. Mégret (UB1)
V13	02/07/2012	Review and finalize the audio related parts	A. Sorin (IBM), A. Satt (IBM)
V14	03/07/2012	Comments, minor modifications	Ceyhun Burak Akgul (VISTEK ISRA VISION)
V15	03/07/2012	Text modifications	Carlos Fernando-Crispim Junior (INRIA)
V16	9/07/2012	Minor modifications	Ioannis Kompatsiaris (CERTH)

			Athina Kokonozi (CERTH)

Executive Summary

The aim of this deliverable is to describe the collection and manual annotation of sensor data for training purposes to support the analysis and interpretation tasks of WP3 and WP4.

Initially the concept and the goals of the Dem@Care project are stated. The next section describes the specifications, the potential functionalities and constrains of sensor devices planned to be used for gathering data in lab environment for Dem@Care, followed by the presentation of objectives and challenges of the annotation and the visualisation tools while the training process is also explained. An overview of existing tools for annotation and visualisation is provided along with the corresponding potential functionalities and constraints according the technical and clinical needs of the project. The specific needs of annotated data required for training and performance evaluation process of Dem@care sensing components for each sensor are also described.

The last section describes benchmarking datasets available for each sensor (or multi-sensor platform), and potential interests for the Dem@Care project while plans for the use of each sensor during the pilots of the project are presented.

Abbreviations and Acronyms

POV	Point Of View
IADL	Instrumental Activity of Daily Living
ADL	Activities of Daily Living
SUP	Scene Understanding Platform
WIMU	Wearable Inertial Monitoring Unit
NTP	Network Time Protocol
MMSE	Mini Mental Score Exam
HRI	Human Robot Interaction
FAB	Frontal Assessment Battery
AI	Apathy Inventory
UPDRS	Unified Parkinson's Disease Rating Scale ()
iVi	Interactive Video Indexer

Table of Contents

List of Figures

List of Tables

1 Introduction

Deliverable D2.3 “Training Data, Collection and Annotation” is the third deliverable in WP2 (Requirements and Impact). Data collection and annotation is supported by clinical experts, as they are responsible to run the Clinical Protocol Pilots, to indicate which measurements are clinically relevant, and to manually annotate the initial training dataset.

Technical Partners responsible for the development of sensing components are involved in the specification of the data collection and annotations processes, as well as in monitoring the data quality during the collection processes (in terms of acquisition and annotation).

In that sense, the report aims at defining the specifications of sensors to be used during the data acquisition process in a controlled lab environment, and the requirements in terms of manual data annotation. Annotated data are used for both training purposes to support the analysis and interpretation tasks of WP3: Health and Lifestyle monitoring and Analysis and WP4: Situational Analysis of Daily Activities, and performance evaluation process of the Dem@Care sensing components.

The primary objective of this report is to provide details of sensors used for data acquisition and to give an overview of the existing annotation and visualisation tools, summarizing their potential functionalities and constraints in the context of the Dem@Care project. The report also aims to specify the training process requirements, and to identify data that can to be manually annotated in the initial training dataset. Finally, this deliverable aims to describe existing benchmark data sets which could also be used for the evaluation of Dem@care sensing components.

2 Concept and Goals of Dem@Care

Dem@Care concerns the development of an integrated solution for the remote monitoring, diagnosis and support of people with dementia. More specifically, Dem@Care aims to enhance diagnosis and to afford timely, personalized, intervention and adaptation of the care being provided. To address the complexity and heterogeneity of the disease and to accomplish the goals, Dem@Care will investigate the use of multiple sensors, for the recording, among others, of daily activities, lifestyle patterns, emotions, speech impediments, and vital signs, as well as the use of intelligent mechanisms for the assessment of the individual's condition and of the appropriate care treatment.

Clinical evaluation and validation will be performed via a three staged evaluation process in pilots that will be carried out in Ireland, France and Sweden, in collaboration with regional clinics, residential care centres and health councils. Objectives of the project and measures of its success are:

- The improvement of the quality of life of people with dementia,
- The advancement of clinical research, by correlating behavioural and cognitive monitoring parameters with dementia-specific patterns,
- The increase of the number of cases with accurate and earlier diagnosis of dementia (especially for patients with Mild Cognitive Impairment)
- The prolongation of time that people with dementia can stay at home, delaying or making unnecessary the admittance to specialized nursing centers (for already diagnosed patients).

3 Sensors & Data Acquisition

This section describes specifications of each sensor to be used during the data acquisition process in Dem@Care pilots. These specifications explain the potential such as the type of information that can be acquired, the functionalities and constraints of the sensor such as the acquisition software with its acquisition frequency and accuracy, the storage software with its storage and transmission protocol, in accordance with the context (e.g. embedded in a lab or at home, at night, with multi-users) and purpose of use (e.g. diagnosis, interaction) in the framework of the Dem@Care project.

3.1 Ambient 2D video camera

3.1.1 Sensor description

3.1.1.1 Context

Ambient 2D video camera point of view (POV) provides a global description of the activity undertaken by the participant and the ability to follow their global trajectory during the recording session in the scene. An example of such off-line processing has been developed by INRIA-STARS team called Scene Understanding Platform (SUP)[1]. Using 2D video stream enables to detect and track people in the scene automatically, to measure the time spent by a participant to perform a predefined IADL (e.g., Coffee preparation), to extract kinematic parameters (e.g., walking speed, distance travelled and trajectory done inside the room) of the participant.

3.1.1.2 Sensor configuration

The 2D video cameras are placed in the room in order to capture all activities undertaken by the participants, and to limit occlusion problems. Computer used for 2D video recordings are synchronized using Network Time Protocol (NTP) to allow further synchronization with other sensors.

[1] http://team.inria.fr/stars/software/

3.1.2 Functional specifications

3.1.2.1 Data recording

Table 1: Data recording specifications for ambient 2D video camera

Model of Video Camera	Fixed network Camera. (Axis® P13 Series).
Type of data recorded	2D images - RGB images. (Compression format: M-JPEG files format; VGA resolution: configurable: 640×480 pixels).
Timestamp availability	Timestamp available. (Timestamp used is the one of the computer running the job).
Data Recording Mechanism	Stored on a server via Internet Protocol (IP) Camera (Ethernet or Wireless).
Data Logging Frequency	Configurable: approx 8 images/sec.

3.1.2.2 Power Source and Internet Protocol video transmission

Table 2: Power Source and Internal Protocol video transmission specifications, for ambient 2D video camera.

Supply Type	Axis P13 Series camera uses external power source. Possibility to use another video camera models using “Power over Ethernet” (PoE) for powering cameras.
Battery lifetime	No battery.
Internet Protocol video transmission	Ethernet or Wireless.

3.1.2.3 System requirements for data recording

Table 3: System requirements for data recording for ambient 2D video camera.

Hardware	No specific requirements.

3.1.2.4 System requirements for data processing

Table 4: System requirements for data processing for ambient 2D video camera

Software for processing to be extended within WP4	• 3D Reconstruction tool to perform the calibration of the camera and to model the scene viewed by this camera (definition of zones of interests, contextual static objects of the scene), • Computer Vision Platform for processing video recordings. (SUP developed by INRIA-STARS Team). • ScRek software (developed by INRIA-STARS Team) for event recognition using muti-sensor approach.

3.1.2.5 Data processing mechanism

During capture, the video is stored on a server via Internet Protocol Camera. Processing is then done offline, after uploading the data to the processing station, using the tools to be developed by the Dem@Care partners within WP4, in order to provide information about the global trajectory of the person tracked in the scene and to perform activity analysis and recognition (activity detection, kinematic parameters extraction of person tracked).

3.1.3 Potential functionalities and Constraints

Table 5: Potential functionalities and constrains for ambient 2D video camera

Position	Occlusion problems due to the POV of ambient video camera.
Comfort	No-worn sensor.
Data Recordings	Enable to clearly identify activities (when used to provide ground truth) undertaken by the person tracked (activities semantics).-RGB images.
Diurnal pattern	Un-usable under adverse lighting conditions (such as sudden changes in the illumination level) because of the deterioration of Computer Vision Platform performance due to illumination changes.

3.2 Wearable video camera

3.2.1 Sensor description

3.2.1.1 Context

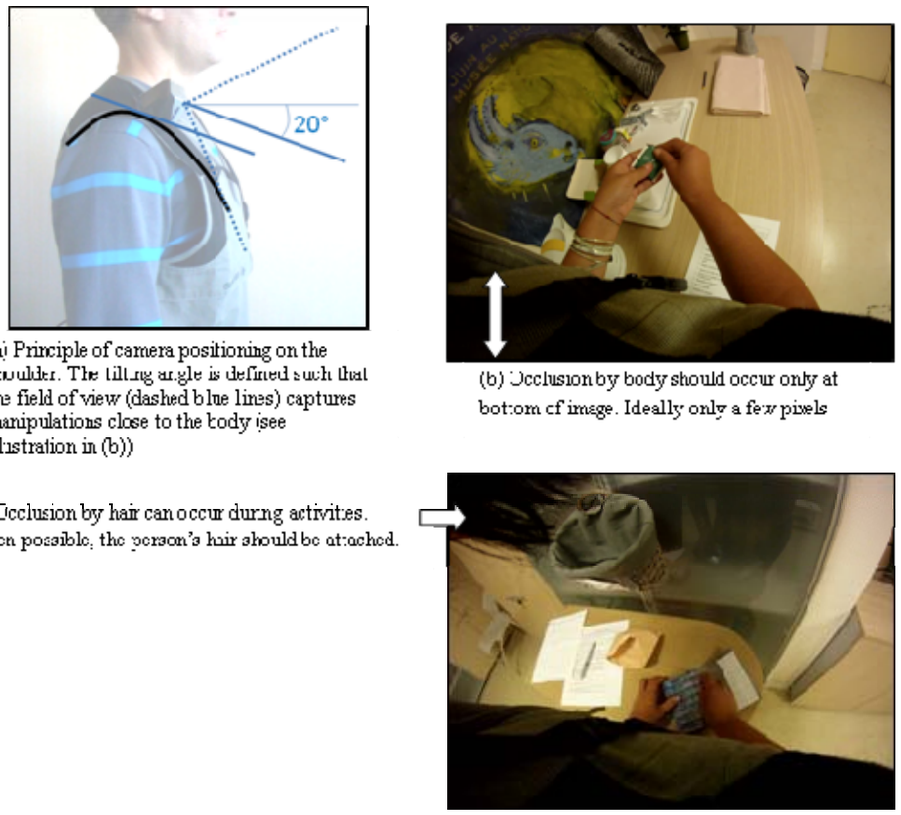
The wearable video camera provides a privileged view for IADL capture. The system developed by UB1/IMS is designed to be worn on the shoulder. It shows a contextual view of the activities of the person, by capturing the interactions with objects from a POV that is very close to the person's subjective view, as well as by providing a wide angle view of the context in front of the person, which characterizes the location of the person and the environment. It can be used in any context where recording of the person instrumental actions, activities and interaction with the environment is useful.

3.2.1.2 Sensor configuration

The camera is fixed on a lightweight jacket worn by the person thanks to hook-and-loop fasteners. It has to be positioned such that the field of view captures the instrumental activity zone in front of the person, without being occluded by the shoulder. A specific foam support has been designed for that matter, to tilt the camera by an appropriate amount, when fixed on the shoulder. Once adapted to a specific person, the camera does not need to be removed from the jacket.

Figure 1 shows the camera on its foam support, and fixed on the jacket (image on the right). It also illustrates the tuning of the field of view in order to optimize the observation of instrumental activities (image on the left).

Figure 1: GoPro camera



(a) Principle of camera positioning on the shoulder. The tilting angle is defined such that the field of view (dashed blue lines) captures manipulations close to the body (see illustration in (b))

(b) Occlusion by body should occur only at bottom of image. Ideally only a few pixels

(c) Occlusion by hair can occur during activites. When possible, the person's hair should be attached.

Figure 2: Tuning of the wearable camera field of view

The wearable camera system is based on a GoPro HD camera. Several modes of acquisition can be programmed: the spatial and time resolution is specified by a mode. The GoPro HD camera provides pre-established configuration modes. We are going to use the video mode r4 which provides 4:3 HD Video with the max Overall View [19].

3.2.2 Functional specifications

3.2.2.1 Data Recording

Table 6: Data recording specifications for the wearable video camera.

Type of data recorded	• Video sequence (H.264 compression, 1280x960p@30im/s in mode 4, 170° angle of view). • Audio (48kHz, mono from lateral microphone, AAC compression)
Timestamp availability	Timestamp available.
Data Recording Mechanism	Stored on removable SD-card. Recording capacity of up to 5h in mode 4 with 32Go SD-card.
Data Logging Frequency	30 images/sec for video.

3.2.2.2 Power Source

Table 7: Power Source specifications for the wearable video camera.

Supply Type	Integrated swappable rechargeable battery.
Battery lifetime	Single charge lasts max 2h30 of video recording. Can be doubled with battery extension that fixes at the back of the camera.
Replacement	USB charger.

3.2.2.3 System requirements for data processing

Table 8: System requirements for data processing for the wearable video camera.

Software	H.264 decoding libraries, such as ffmpeg, VLC or avidemux.

3.2.2.4 Data processing mechanism

During capture, the video is stored on the onboard SD-card. At the end of the recording session, the SD-card is plugged into a collecting computer, which downloads the H.264 video files. Processing is then done offline, after uploading the data to the processing station, using

the tools to be developed in WP4, in order to provide information about object detection, localization, and activity analysis.

3.2.3 Potential functionalities and Constraints

Table 9: Potential functionalities and Constrains for the wearable video camera

Position	Fixed on the shoulder with hook-and-loop fasteners to a dedicated lightweight jacket.
Comfort	• Can be seen by the other people and so may cause some social awkwardness. • Requires to put the specific jacket, but becomes unnoticed to the wearer after starting activities.
User interaction	• Must download images from the SD-card and charge device after 2h30 to 5h of recording.
Diurnal pattern	• Can be worn for most activities when standing or sitting. Image quality can be poor in badly-lit environments, or where wearer is moving rapidly. For privacy concerns the stop button is accessible, but specific arrangement may be required if the device is used without any external assistance.

3.3 Wearable still image camera (SenseCam)

3.3.1 Sensor description

3.3.1.1 Context

The SenseCam has been used in many different life-logging applications. Of particular interest to Dem@Care are applications such as day-to-day photo logs of wearer's POV. These logs may be used for reminiscence therapy and also for automatic activity detection and classification. Generally speaking the device can be used in any context where recording of the wearer's POV is useful.

3.3.1.2 Sensor configuration

SenseCam is a small digital camera that is designed to take photographs automatically without user intervention. SenseCam contains a number of different electronic sensors that can be used to collect data for the lifelogs: light-intensity and light-colour sensors, a passive infrared (body heat) detector, a temperature sensor, and a multiple-axis accelerometer. These sensors are monitored by the camera's microcontroller, and changes in sensor readings can be used to automatically trigger the camera shutter. In addition to image data, the memory card stores a log file, which contains the sensor readings data at each time that an image is taken, the reasons for taking each image (e.g. manual shutter press, timed capture or significant change in sensor readings), and timestamp information.

3.3.2 Functional specifications

3.3.2.1 Data Recording

Table 10: Data recording specifications for SenseCam.

Type of data recorded	• Photographic images. • GPS data. • Accelerometer data.
Timestamp availability	• Timestamp available.
Data Recording Mechanism	Stored on device and downloaded to repository on approximately daily basis (depending on current application usage).
Data Logging Frequency	• Configurable: approx 1 image / minute in normal usage • Image is recorded when light-change sensor is triggered (e.g., move from indoors to outdoors).

3.3.2.2 Power Source

Table 11: Power Source specifications for SenseCam.

Supply Type	In-built rechargeable battery.

Battery lifetime	Single charge lasts approx 16 hours of full usage.
Replacement	USB charging.

3.3.3 Potential functionalities and Constraints

Table 12: Potential functionalities and Constrains for SenseCam.

Position	Worn on lanyard around the neck, hanging around chest-height.
Comfort	• Can be obvious and so may cause some social awkwardness. • Should be generally comfortable from physical POV.
User interaction	Must plug in USB to download images and charge device on daily basis.
Diurnal pattern	• Can be worn for most activities but image quality can be poor in badly-lit environments, or where wearer is moving rapidly. For privacy concerns there is a button which pauses operation for a short interval.

3.4 3D video camera: Kinect® RGBD

3.4.1 Sensor description

3.4.1.1 Context

Kinect® is a 3D video camera developed by Microsoft Corporation. It provides a 2-D RGB image of the scene associated with a depth map estimation built using infrared sensors. The combination of these sensors produced an image estimation referred as RGB-D. Besides Microsoft Kinect SDK, a few free and open-source libraries are available as an alternative to support the development of applications for Kinect. This sensor can be also used for ambient video recording as the ambient 2D video camera, but it is generally recommended to the

detection of body part movements and accurate body movement (e.g., for gesture recognition), both performed closer to the sensor (up to 5 meters).

3.4.1.2 Sensor configuration

The Kinect® is connected to the computer via USB cabling. Before recording session, it should be initialized in order to provide a proper 3D-estimation of the scene. Previous projects of INRIA-STARS team have used the libfreenect software for this purpose.

3.4.2 Functional specifications

3.4.2.1 Data Recording

Table 13: Data recording specifications for Kinect®.

Type of data recorded	For one frame, 3 types of data are stored: • 2D image (RGB image). • Depth map. • Intensity data.
Timestamp availability	Timestamp available (timestamp used is the one of the computer running the job).
Data Recording Mechanism	Data (images, depth map, intensity) are directly stored on the computer connected to Kinect Sensor via USB port.
Data Logging Frequency	Approx 10 frames/sec.
Storage Volume	Approx 3Mb/frame.

3.4.2.2 Power Source

Table 14: Power Source specifications for Kinect®.

Supply Type	Kinect sensor has to be plugged into an external power source and then plug it into the computer's USB port.
Battery lifetime	No battery, use external power source.

3.4.2.3 System requirements for data recordings

Table 15: Data recording requirements for Kinect®.

Hardware	• 32 bit (x86) or 64 bit (x64) processor. • Dual-core 2.66-GHz or faster processor. • Dedicated USB 2.0 bus. • 2 GB RAM.
Software	• libfreenect software: recordings could be done using Kinect drivers with OS Windows/Linux/ Mac OS. • Possibility to record using plugin developed on the Computer Vision Platform (software developed by INRIA- STARS Team).
Material	Special USB and power cabling.

3.4.2.4 System requirements for data processing

Table 16: Data processing requirements for Kinect®.

Software	• 3D Reconstruction tool to model the scene viewed by this camera (definition of zones of interests, contextual static objects of the scene). • Computer Vision Platform for processing video stream.

3.4.2.5 Data processing mechanism

Data processing mechanism for the Kinect sensor is the same than this one described in the part 3.1.2.5.

3.4.3 Potential functionalities and Constraints

Table 17: Potential functionalities and Constrains for Kinect®.

Position	• Scope limitation of the Kinect to record depth map: ~5meters. • For good results the person tracked has to be at a distance between two and five meters from the Kinect sensor.

Comfort	No-worn sensor.
Data Recordings	• Enable to clearly identify activities (ground truth) undertaken by the person tracked (activities semantics) (RGB images). • Assignment of a skeleton to each person (articulation) from the Kinect sensor's POV. This data significantly improve the automatic people detection/tracking in Computer Vision, and could also be used as ground truth for People detection/tracking. • Body part detection enabling a good description of person's gesture. • Potential for audio recordings.
Data Storage	• Volume of data stored if data processing is not done on line (approx.: 3Mb/frame).
Diurnal pattern	Not sensitive to light-change due to infrared sensor.

3.5 Accelerometers

3.5.1 Sensor description

3.5.1.1 Context

The Wireless Inertial Measurement Unit WIMU devices have generally been used in measuring athletic performance, by recording information such as orientation, impact, movement, direction, etc.

In the context of Dem@Care, it is expected that these devices will be used to monitor motion in general, but also for specific applications such as gait analysis. The devices will also provide some actigraphic information that could be used for activity classification. Furthermore, the accelerometers in these devices could be used to provide a fall detection functionality.

3.5.1.2 Sensor configuration

The sensors need to be calibrated before data logging is performed.

3.5.2 Functional specifications

3.5.2.1 Data Recording

Table 18: Data recording specifications for WIMU.

Type of data recorded	Motion data: • Accelerometer data. • Gyroscope data. • Digital compass .
Data Recording Mechanism	Data is transmitted wirelessly to a base unit connected via USB to a computer which performs the data logging.
Data Logging Frequency	• Configurable. Logging can be tailored to suit system requirements. • Trade-off between amount of data logged and battery lifetime.

3.5.2.2 Power Source

Table 19: Power Source specifications for WIMU.

Supply Type	WIMU devices are charged via USB.
Battery lifetime	Approximately 1 hour at 5Hz logging rate.

3.5.3 Potential functionalities and Constraints

Table 20: Potential functionalities and Constrains for WIMU.

Position	• Wearable: e.g. wrist, neck, waist. • Matchbox size design means the WIMU can be worn on most parts of the body, with minimal intrusion. (Has been previously used with High-Performance athletes with no intrusion on their motion, positioning or actions.).
Diurnal pattern	Can be worn for most activities (but may vary with activity and measurement to be undertaken).
Reliability Consistency	Devices must be calibrated before use to ensure consistent data.

3.6 Audio sensor

3.6.1 Sensor description

3.6.1.1 Context

The audio sensors, namely microphones, are used for collecting high-quality voice sounds, from people being diagnosed and from the clinicians who diagnose them. The recorded audio is utilized for developing vocal bio-markers and classification algorithms, which aim at supporting early detection of dementia, tracking dementia state from remote, identifying certain emotional states and mood of people with dementia, and therefore supporting the core purposes of the project. The audio sensor system consists of two wearable wireless microphones – one to be attached to the person under diagnosis and the other to the clinician, an adaptor unit, and a general purpose computer to record the sounds signals into sound files. Ambient audio microphone is used as well, as a secondary source of audio.

3.6.1.2 Sensor configuration

The two audio streams will be stored in a stereo audio file with the participant's stream as the left channel always. The wearable microphone recording will be done at 22050 Hz, 16 bits per sample per channel.

3.6.2 Functional specifications

3.6.2.1 Data Recording

Table 21: Data recording specifications for audio sensors.

Type of data recorded	Stereo audio file (×2, 1 for Participant+1 for Clinician).
Timestamp availability	Manually.
Data Recording Mechanism	Audio data is transmitted wirelessly from the microphone to a base unit connected to a computer.
Data Logging Frequency	Frequency: 22050 Hz, 16bits per channel.
Software	Using any general purpose audio software (the software to be defined during the project).

3.6.2.2 Power Source

Table 22: Power Source specifications for audio sensors.

Supply Type	Regular 1.5V batteries of size AA.
Battery lifetime	About 4-5 hours.
Replacement	Regular 1.5V batteries of size AA.

3.6.3 Potential functionalities and Constraints

Table 23: Potential functionalities and constrains for audio sensors.

Position	Worn next to the mouth (max. Distance wearable microphone/mouth= 20 cm).
Comfort	Can be seen by the other people and so may cause some social awkwardness.
Reliability / Consistency	No issues to refer.

3.7 Physiological sensors

3.7.1 Sensor description

3.7.1.1 Context

The BodyMedia SenseWear Pro3 and Philips DTI-2 are devices suitable for conducting research on different aspects of physiology and lifestyle monitoring. Both devices allow logging of a number of different parameters: galvanic skin response (GSR), accelerometers (for capturing motion), temperature and ambient light (only for DTI-2). In principle, the DTI-2 devices provide a better technical specification than the BodyMedia devices: smaller form factor, 3D instead of 2D accelerometer, better wearing position for measuring GSR, more storage capacity, etc. Due to limited availability of the DTI-2 devices and unknown reliability in the field, the BodyMedia devices are therefore seen as a fallback option.

3.7.1.2 Sensor configuration

Both the BodyMedia devices and the Philips devices need to be configured once. For the BodyMedia devices this can be done through the proprietary BodyMedia software. For the Philips devices a simple text configuration file can be edited.

3.7.2 Functional specifications

Table 24: Functional Specifications for BodyMedia SenseWear Pro3 and Philips DTI-2 devices.

Devices	BodyMedia SenseWear Pro3	Philips DTI-2
Type of data recorded	• Skin conductance, 2D accelerometry, temperature. • Proprietary data format.	• Skin conductance, 3D accelerometry, temperature, ambient light. • ASCII text columns.
Timestamp availability	Timestamp included in data.	Timestamp included in data.
Data Recording Mechanism	• Internal memory. • Offloading via USB to PC using proprietary software.	• Micro SD card. • Connects as USB mass storage device.
Data Logging Frequency	Sample rate configurable.	Sample rate configurable (typically fs=2-10 Hz).
Supply Type	Battery operated.	Battery operated.
Battery lifetime	Up to one week (storage is limiting factor).	Up to one week.

3.7.3 Potential functionalities and Constraints

Table 25: Potential and constrains for BodyMedia SenseWear Pro3 and Philips DTI-2 devices.

Devices	BodyMedia SenseWear Pro3	Philips DTI-2
Position	Upper arm.	Wrist worn.
Comfort	Relatively comfortable, device tends to slip down during lengthy measurements.	Unobtrusive, device is worn as wrist watch.
User interaction	No interaction required.	No interaction required.
Diurnal pattern	Can be worn for most activities, not for use in/under water	Can be worn for most activities, not for use in/under water.
Reliability/co	Devices have been proven to have	Relability is unknown; new

nsistency	reliability issues: Recording only starts when device detects skin contact, which doesn't always occur. Sometimes measurement sessions are completely/partially lost.	research prototype.

3.8 Data acquisition constraints for a multi-modal sensor analysis

The use of several sensors could improve the participant's activity analysis but constraints about data acquisition must be taken in account for optimal data synchronization. Besides to common data storage format readable, the readings of sensor must be provided with a timestamp. From previous projects involving clinical protocols similar to Dem@Care, the sensor acquisition rate of 1reading/sec is not enough to capture short time activities.

We envisage that the acquisition rate must be equal or higher than 4readings/sec to fully capture the activities characteristics (e.g., gait parameters, posture transfer).

4 Annotation tools

In this section, we first present the objectives and challenges of annotation process through the development of Dem@Care sensing components. Then an overview of existing tools currently used for video annotation is provided, as well as potential functionalities and constraints of these tools in accordance with the annotated data required for training process.

4.1 Objectives and Challenges

The objective of an annotation tool is to support experts at describing the meaning of raw sensor data. Although the Dem@Care project deals with multi-sensor approach, video-based annotations are going to be used as reference to determine human activity instances in respect to different type of sensor readings. In this perspective, optimal software for video-based annotation should address the following functionalities:

- Ability to zoom on the image: if the quality of the video is bad, it is not easy to determine where the object of interest is, zooming on the image will enable the user to better determine the contour of the objects of interest.
- Bounding box orientation: depending on the camera POV, the objects of interest can be leaned on the video, as such an oriented bounding box (or ellipse) is necessary to accurately locate the object.
- Interpolation: to avoid the user drawing bounding box around object of interest, interpolation from frame n to frame n+t should be done. The best interpolation should be based on image characteristics and not only on the size of the bounding box.
- Attribute personalization: the tool should enable the user to personalize the attributes associated to the annotated object (posture, gender, role).
- Event annotation: the interface should provide an easy way to annotate activities happen in the video.

Moreover, if the experiments are performed with several sensors, correspondence (i.e., spatial coherence of the object of interest across sensors used) and synchronization (i.e., temporal coherence of event time period across sensor used) should be considered. If both these

constraints are respected, annotation for one sensor can be automatically translated for the other sensors, respecting each sensor capabilities.

4.2 Annotation tool description

4.2.1 Annotation tool for processing ambient 2D video camera data

4.2.1.1 ViPER

ViPER is a set of tools proposed by University of Maryland for evaluation purposes. It consists of ViPER-PE (for performance evaluation) and ViPER-GT (for ground truth annotation). In this part, we will focus on the second tool. Description of the tool and documentation explaining its use is available[2].

The Video Performance Evaluation Resource Kit's Ground Truth tool, or ViPER-GT, allows someone to annotate a video with metadata, mainly for use as ground truth for performance evaluation. This includes general information describing the file, such as date of filming and keywords about its content. It also includes concrete features, such as scene breaks and bounding boxes around people. This can be used for any number of purposes.

[2] http://viper-toolkit.sourceforge.net/docs/

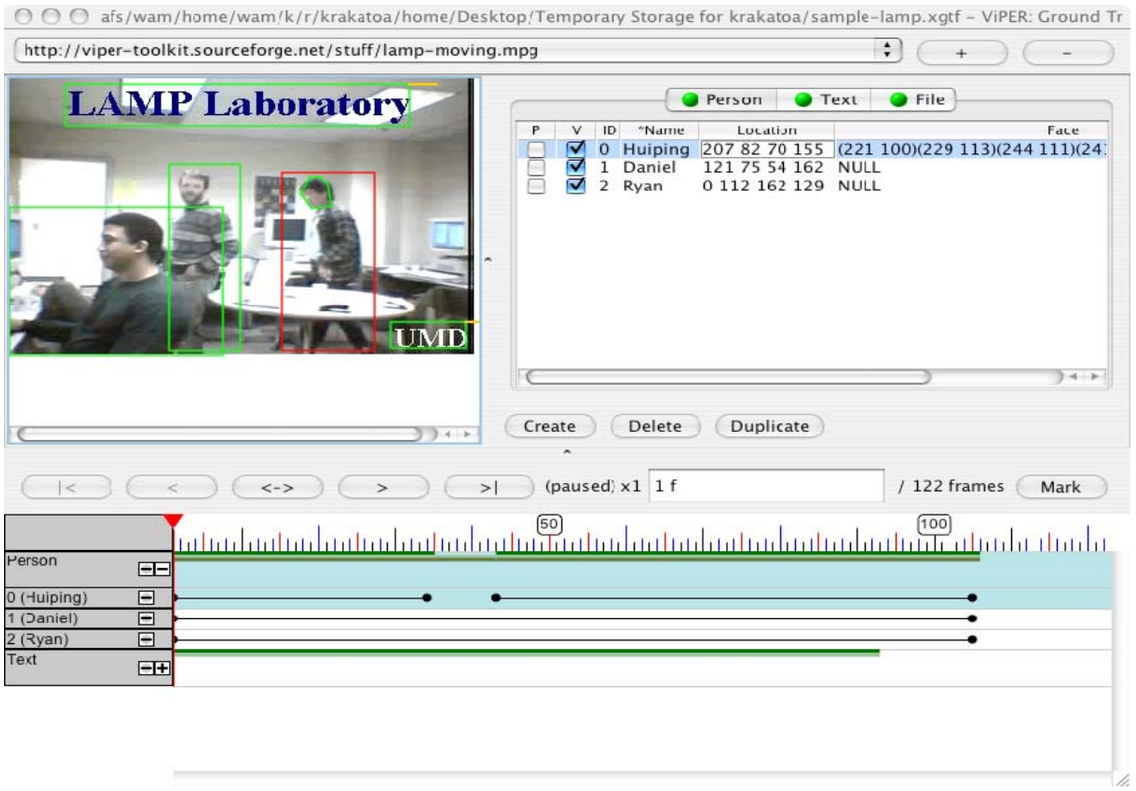


Figure 3: ViPER-GT Interface

ViPER-GT is composed of three main parts (Figure 3). The video frame pane is interactive in that the user can manually draw bounding boxes for instance. On the right of the video frame pane is the spreadsheet view. Across the top are tabs for each descriptor type. In the sample video, these are 'Text', 'Person', and 'File'. Each table has a row for each instance of the descriptor type. One can edit all the values directly in the table, but some of the values may also be edited on the video frame view or the timeline view. Below the frame and spreadsheet views is the timeline view. It displays a summary of the entire video. There is a summary line for each type of descriptor, which can be expanded to show when each descriptor of that type is declared to be valid. It is also possible to use the red arrow marker to scrub the video.

4.2.1.1.1 Functionalities

The ViPER-GT tool allows to manually annotate a video in terms of object of interest (e.g. person and associated characteristics: posture, identity, etc.), and in terms of events (e.g. activities associated to a person in a video sequence).

The tool is designed for editing visual annotation, such as rectangles denoting locations of people on screen. The current shapes in version 4.0 include points, bounding boxes and oriented rectangles, ellipses, polygons and circles. There are also types without a visual element, including text strings, numbers, and boolean values.

Data elements are combined together into objects called descriptors. This allows defining a person type, which has a text string (the person's name), a bounding box (their location), and any other number of attributes. Descriptors usually refer to a single object, event or other thing in the file that is worthy of evaluation, but they may also have more abstract purposes, such as indicating key frames. In addition, all files have a single descriptor that gives metadata about the media file as a whole, including frame rate, file name, image size in pixels, and an optional comment.

ViPER-GT maintains a set of descriptors associated to various source media files. You can have one annotation file that describes several different media files, although it is often useful to have a one-to-one mapping of media file to annotation file. It also presents a schema editor for describing what kinds of descriptors you may mark up.

Finally, ViPER-GT allows users to perform two operations on values of attributes of selected object descriptors over a range of frames. The operations are propagation and interpolation. It is possible to invoke these operations directly from the table by right-clicking on the descriptor, or through the timeline by dragging from one place to another.

4.2.1.1.2 Tool characteristics

The tool is available in two formats: the lite one and the extended one. It can run on multiple platforms as the code is available as Java classes. The lite version contains a .jar file. The extended package requires tcsh or sh shell, Java2, and Perl.

ViPER-GT takes as input a mpeg file or a list of JPEG files, and output the manual annotations as a XML file.

ViPER-GT has been widely used in different projects as metro surveillance, Alzheimer patient monitoring or airport application. In particular, it is used by ViSEvAl (section 5.2.1) tool developed by STARS team at INRIA for multi sensors activity evaluation purpose. We will describe here the file format for annotation of person and event in video used by ViSEvAl. The format is easily adaptable to the Dem@Care requirements.

The first tag (config) of the XML file describes the configuration for the annotation: what the user want to annotate. Each object of interest is then described. For video application, two kinds of interesting objects are annotated: the physical object and the events. Each object is then described with different features. ViPER-GT can configure the possible values for each attribute to make annotation more easily.

The physical object is described with:

- Type: can be NULL, Person, Vehicle or Equipment
- Subtype: the value will depend of the application (e.g. Nurse, Patient, Chair)
- Info2D: is the bounding box around the annotated object

The event is described with:

- Name: the value will depend of the application (e.g. Walking, Reading, Cooking).

Moreover a frame span (when the object appears or when the event occurs) is associated to each annotation.

Then, general information for the video sequence is described. This information consists of:

- SOURCETYPE: it can be a SEQUENCE or a FRAMES.
- NUMFRAMES: the number of frames in the considered video.
- FRAMERATE: the number of image by second of the video.
- H-FRAME-SIZE: the width of one frame in pixel.
- V-FRAME-SIZE: the height of one frame in pixel.
- FIRST-FRAME: the number of the first annotated frame.
- CAMERA: the name of the camera, useful if several cameras are available for the application.

A short example of this output file is available in APPENDIX A1. The second tag (data) of the XML file describes the different values annotated by the user.

4.2.1.1.3 Potential functionalities and Constraints

ViPER-GT tool is a free tool, which can be adapted to any application field. Indeed user can configure each attribute (posture, dominant color, etc.) associated to the tracked person. It can be used for events annotation and also to detect and track different physical objects in the scene (persons, moving objects such as household appliances).

Among ViPER-GT restrictions are the possibility of annotate only mono-camera video sequences with fixed POV, and the compatibility only with MPEG video format. In our previous experiences, ViPER module for events annotation is considered not user-friendly by the Clinical staff. Users have preferred to use a visualisation tool and a excel spreadsheet file where they directly fill the visualized event information (event name, and the beginning and end frames). The excel files were later translated by technical part into ViPER files.

4.2.1.2 ANVIL

ANVIL is a free video annotation tool[3] shown in **Error! Reference source not found.** It offers multi-layered annotation based on a user-defined coding scheme. During coding, the user can see color-coded elements on multiple tracks in time-alignment. Originally developed for gesture research in 2000, ANVIL is now being used in many research areas including human-computer interaction, linguistics, ethnology, anthropology, psychotherapy, embodied agents, computer animation and oceanography.

Figure 4: The ANVIL interface

[3] http://www.anvil-software.de/

4.2.1.3 Functionalities

ANVIL is used to annotate the scene shown in a video, describing the relevant elements of what is happening. The coding scheme defines *tracks*, representing properties describing the scene. During annotation, track elements are added to the time-line. Figure 5 illustrates the annotation procedure which consists in adding track elements for the predefined tracks. A track element is defined within a start and end time and contains multiple attributes to encode many aspects of one "event". Attributes are defined in the scheme with types and value ranges. During annotation, the interface for elements creation is adapted to these types and ranges, ensuring the conformity to the scheme. Additionally, each track element has a *comment* attribute, which is a free space for any kind of additional information. Several elements in the same track cannot overlap.

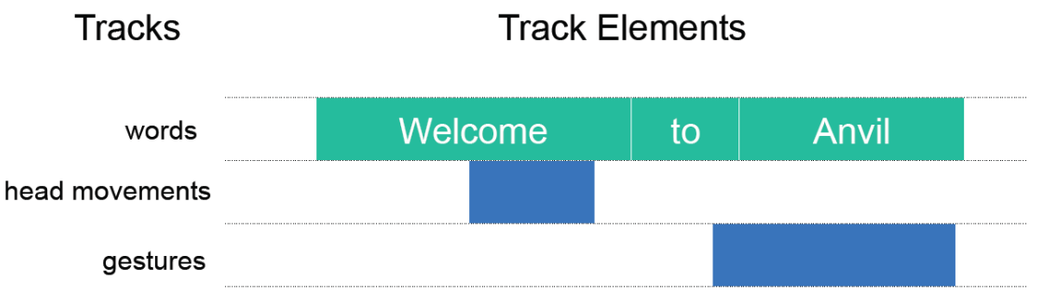


Figure 5: The ANVIL interface (picture from the website)

Figure 4 shows the ANVIL interface, composed of several windows: the main window (top-left) is used for video control, the video window (top-center) shows the video being coded, and the element viewer (top-right) displays the selected element as well as all its attributes and comment. Finally, the annotation board occupies the down window. It contains the record line (horizontal), the playback line (vertical) and the tracks with their elements, the active track being displayed with a different color.

Some special features are cross-level links, non-temporal objects, time point tracks, coding agreement analysis, search of annotated elements and a project tool for managing whole corpora of annotation files.

The synchronous annotation of multiple videos is a functionality offered by ANVIL.

The ANVIL tool incorporates an analysis of the annotations. Analysis creates histograms and other statistics about a track element. It also computes an agreement measure between several

coders. Performed analysis can be exported as a text table, importable to Excel, Statistica and SPSS.

4.2.1.3.1 Tool characteristics

The ANVIL annotation tool is developed in Java, which makes it portable across platforms. It is also possible to develop plug-ins, extending ANVIL with new functionalities. The specification with the coding scheme is defined in a separate XML file. ANVIL offers a graphical interface to create this file. Annotation data is also saved in an XML file, with a reference to the specification and the video files. Several annotation files may be created for a video, to annotate separate semantic elements or to get around the constraint that track elements cannot overlap. For example, to annotate the simultaneous behaviour of several people, different annotation files are necessary.

4.2.1.3.2 Potential functionalities and Constraints

ANVIL is well adapted to describe the rich content of a scene. If the needed annotation should contain many elements about the ongoing activity, then ANVIL is an adapted tool. For purely visual annotation, such as bounding boxes for detecting and/or tracking people or other moving objects, ANVIL can be used as well. The basic version of ANVIL does not include bounding box annotation, but a plug-in exists (an example is shown in Figure 4). Although for bounding box annotation only, simpler tools such as ViPER seem more adapted. Other spatio-temporal annotations may be done using ANVIL, for instance movement paths for annotating gestures.

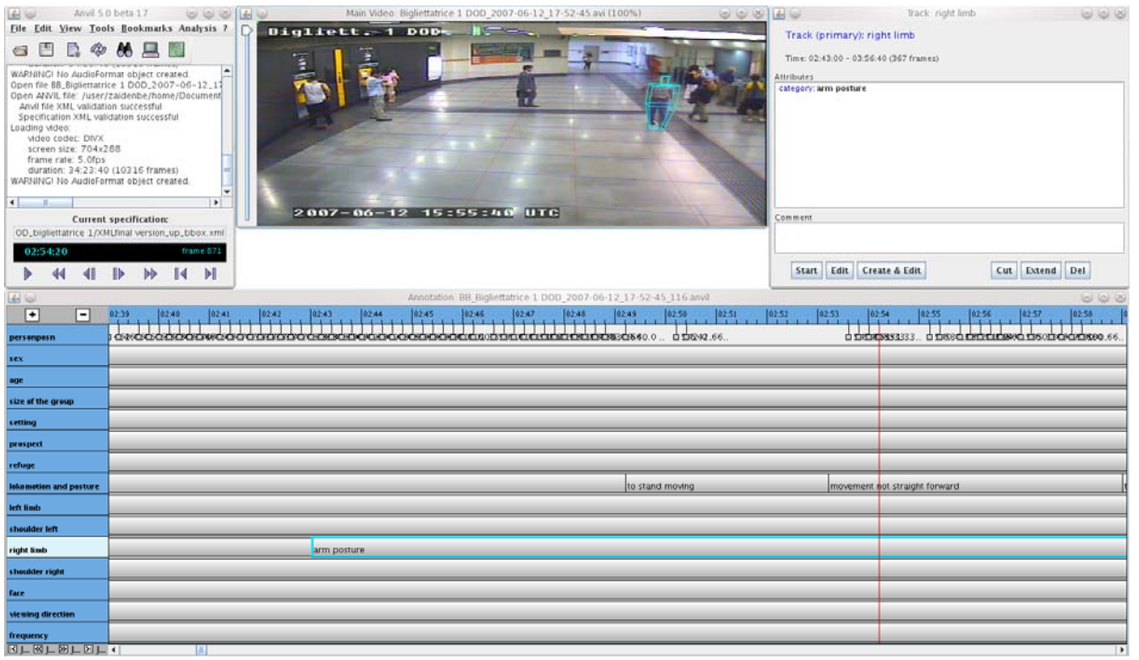


Figure 6: The ANVIL interface, example of the people bounding box plug-in

4.2.1.4 Other existing annotation tools for processing ambient 2D video camera data

In the previous parts, description of annotation tools currently used and/or developed by INRIA-STARS Team was provided. Below, a non-exhaustive list of other ambient 2D video camera-annotation tools is provided.

Table 26: Annotation tools description.

Name	Application field/Site for information
Free Software	
ETHOWARCHER	Annotation tool for behavioural analysis and tracking (single target) [15]
Etholog	Annotation tool for behavioural analysis [16]
Commercial software	
Observer XT	Annotation tool (and visualisation tool) for behavioural Analysis [17]

4.2.2 Annotation tool for processing wearable 2D video camera

4.2.2.1 iVi

iVi (Interactive Video Indexer) is an annotation software developed by UB1/LaBRI. This software has been developed to annotate a video with respect to temporal events, and object occurrences, to constitute a ground truth for performance evaluation.

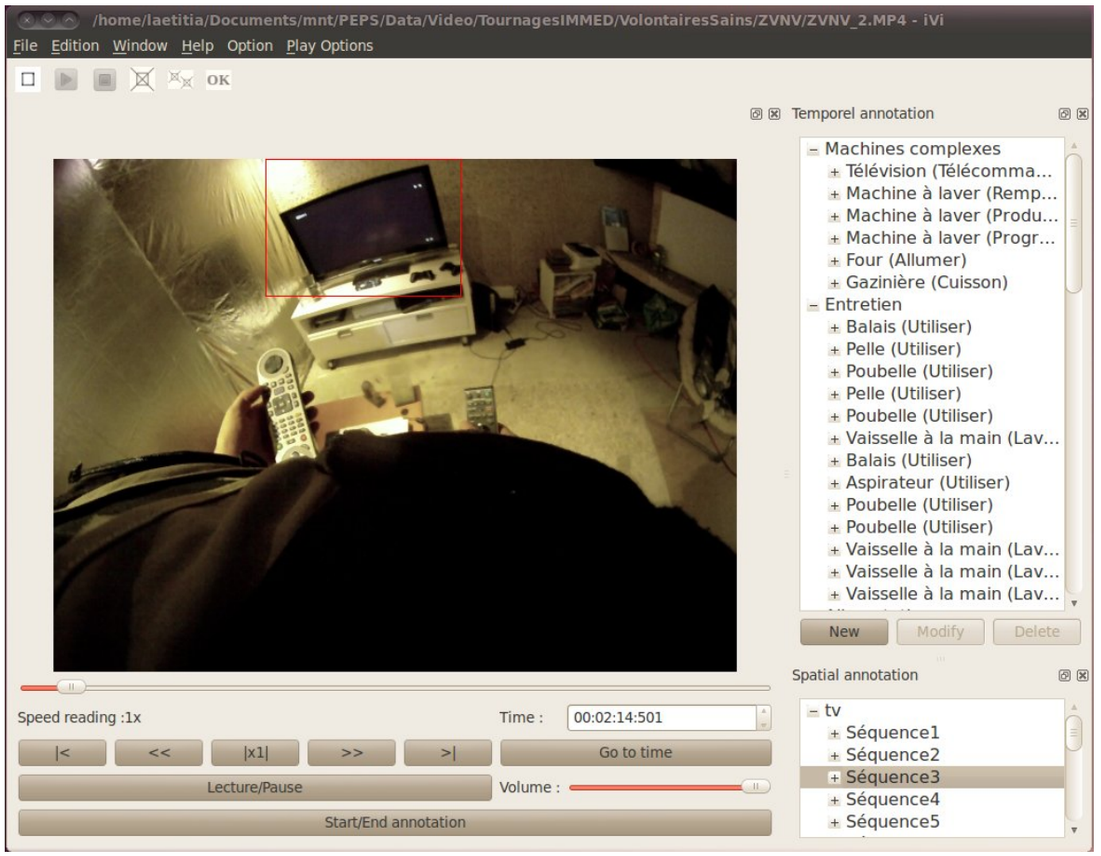


Figure 7: iVi interface

4.2.2.1.1 Functionalities

iVi annotates a video in term of events (activities) or in term of object of interest (objects which are handled by the person during an activity). The iVi interface is composed of four parts (Figure 7). The video frame is interactive and the user can annotate objects by manually defining a bounding box. On the right of the video frame are the annotation views: the first correspond to the temporal annotation (activities), and the second is the spatial annotation (objects of interest). Each view is interactive: the users can modify an annotation, and can

change the start or the end time of the activity for instance. To fill these views, the software has a toolbar for the objects annotation (at the top) and a button “Start/End Annotation” for the activity annotation (at the bottom). Underneath the video, the buttons allow to navigate easily in the video. Thanks to the slider, the user can go directly to an arbitrary time within the video. Moreover, the software gives the possibility to navigate between the annotations based on the events. The user can jump to the previous annotation or the next annotation of the current annotation. Considering the objects annotation, the users can select a bounding box around the object of interest (like the TV in Figure 8). They can interact with the bounding box, modify their size or move their position. An automatic tracking of the object and the manual correction tool are included to facilitate the semi-automatic extension of the manual annotation over a temporal interval.

4.2.2.1.2 Tool characteristics

The iVi software is developed in C++ and uses different libraries such as QT, VLC, OpenCV and FFMPEG.

The software takes as input a video (avi, mpeg, mp4…) and output as a XML file which contains the manual annotations (activities and objects) and the information about the video (name, size, ips). The annotations can be saved into two different formats, which are described in APPENDIX A2.

4.2.2.1.3 Potential functionalities and Constraints

iVi allows to annotate videos to construct a ground truth. With this software, the different activities of the video can be easily and precisely annotated by the user.

4.2.2.2 BBox

BBox has been developed by the UB1/LaBRI to annotate objects in frames (Figure 6). This tool allows annotating a video frame per frame. It can be useful for annotation of small video with no so many frames.

Figure 8: BBox interface

4.2.2.2.1 *Functionalities*

The software is designed to be the simplest tool for the annotation of objects in short sequences. It displays each image to be annotated. An annotation session is started by opening a video and a txt file, which will record the bounding box of this video. Bounding boxes can be created and deleted. Navigation is based on the frame number which is displayed, with the "Next frame" button to move forward in the image sequence.

4.2.2.2.2 Tool characteristics

BBox is developed in C++ and uses the FFMPEG library to decode frames of a video. The software takes as input a video (avi, mpeg, mp4…) and output as a txt file which contains the coordinates of the bounding box of the object. This file is composed of the frame number, the number of bounding box on this frame, and the coordinates of the bounding boxes (x, y, width, height). An example is presented in Figure 9.

0 1 259 501 183 147
1 1 223 734 183 147
2 1 246 742 183 147
3 1 201 774 183 147
4 1 266 774 185 153
5 1 268 791 185 153
6 1 299 799 185 153
7 1 292 795 188 144
8 1 303 787 188 144
9 1 292 813 188 144

Figure 9: BBox file

5 Visualisation tools

In this section, we first present the objectives and challenges of visualisation tools for the performance evaluation process of Dem@Care sensing components and clinical interpretation of multi-sensing data. Secondly an overview of existing tools currently used for visualizing multi-sensing data is provided with its potential functionalities and constraints in accordance with the technical and clinical needs of Dem@Care project.

5.1 Objectives and Challenges

Visualisation is an important step for the evaluation process. The main idea is to be able to display the raw data in a user-friendly way, the associated detections and the annotations. Each kind of sensors has its own raw data output (images for video sensors, value graphics for accelerometers, etc.). Main types of data inputs are:

- Raw data: the first visualisation consists in displaying the output of the sensors to the user (images, graphics, etc.)
- Detection: the detection should be visualized on the raw data. For instance the bounding box of the detected objects can be displayed directly on the images of the video.
- Ground-truth: in the same way than the detection, the annotation will be displayed on the raw data

For Dem@care project, the main challenges of visualisation tool are time synchronisation and spatial correspondence at the visualisation of multi-sensors data input.

5.2 Visualisation tool description

5.2.1 ViSEvAl

At INRIA (STARS Team), an evaluation framework has been developed to assess the performance of Gerontechnologies and Videosurveillance: ViSEvAl (Figure 10). This framework aims at better understanding the added values of new technologies for home-care monitoring and other services. This platform is available to the scientific community and

contains a set of metrics to evaluate automatically the performance of software given some ground-truth. This software has been successfully applied on several kinds of application:

- Airport: evaluates event recognition and video algorithms for activities around aircraft (multi video cameras configuration)
- Metro: evaluates group behaviour and person counting (mono camera configuration)
- Hospital: evaluates Alzheimer patient activities (mono camera and accelerometer)
- Road: evaluates wrong vehicle direction (mono camera configuration)

5.2.1.1 Functionalities

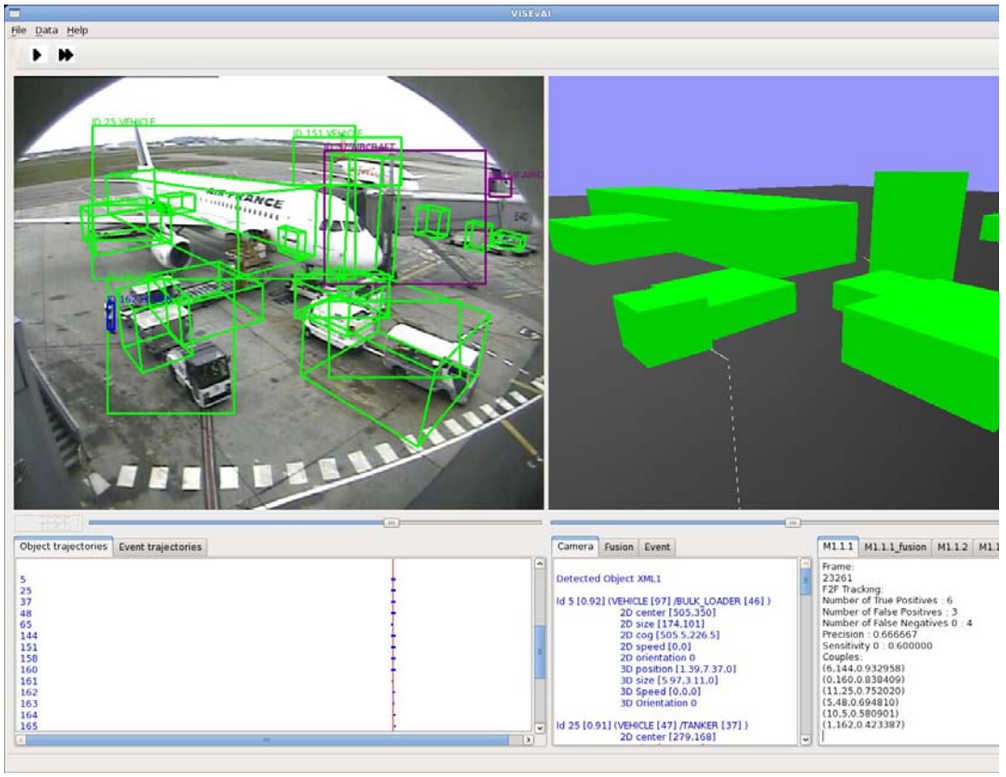


Figure 10: The ViSEvAl software interface

The ViSEvAl (visualisation and evaluation) software aims to propose two main functionalities:

1. Visualisation: support the visualisation of multi-source of information: raw sensor data, results of the algorithms, annotated ground-truth and evaluation results,
2. Evaluation: run evaluation metrics, and easily update the evaluation criteria.

In this document, we will focus on the visualisation part. The visualisation of the different data is possible due to different functionalities:

- Enhanced real video stream: the video stream is displayed, it can be read, stopped, accelerated. The image is enhanced with the object detected by video algorithms, or/and annotated in the ground truth. The objects are represented with bounding boxes with different colours according to their types (e.g. person or vehicle). Moreover, recognised events are also displayed on the image.
- 3D view: the 3D representation of the scene available helps the user to spatially visualize the different detected/annotated objects. The objects can be displayed in different mode: full, wireframe, transparent.
- Temporal information: the information about results and ground truth can be temporally visualized. Different windows are available for the detection and the events. For instance, the user can visualize all the detections for the whole video in one place.
- Evaluation visualisation: curves representing performance metric such as precision and sensitivity are available.

The software synchronises in time all the input data (detection from one or several sensors, ground truth).

5.2.1.2 Tool characteristics

The tool is open source under A-GPL license. INRIA chooses an open source licence because scientific community can share own evaluation criteria, and then enhance the software. ViSEvAl is developed in C++ and dependent of QT4, and xsdcxx. The tool is proposed for Linux system, but porting to Windows system is possible.

The input file type is xml. The description of the different file format is given in the tool with xsd files. We can list the most important files format below:

- XML1 : describes detections obtained from one camera.
- XML2 : describes fused detections obtained from several cameras.
- XML3 : describes recognised events.
- XML5 : describes detection from accelerometer.

Description of all these formats can be found at [18].

ViSEvAl enables to synchronize different output sensors (e.g., video sequence and accelerometer data) and visualize them simultaneously. It is also possible to synchronize and visualize multi-camera sequences simultaneously. This software is Open Source (AGPL resource) and can be upgraded for specific purposes. Many functionalities of the software are implemented with plugins (e.g., video loading) so it can be specialized for given applications. This software can be also used for evaluation purposes with the possibility to compute metrics for different tasks (detection, tracking, events).

ViSEvAl is well adapted for visualisation of data containing at least one video stream. Input files (such as XML files containing detection, tracking and events detected, and XML file for ground truth data) are not flexible. These input files need to be converted to the ViSEvAl format. For the moment, no functionality is associated to audio sensor data.

5.2.2 Consultation iVi

The "*Consultation iVi*" interface, developed by UB1/LaBRI allows visualizing a video thematically from an automatically or manually indexed audio-visual content. This tool provides event based navigation, where the user can visualize an overview of the indexed activities (represented as annotated temporal intervals), and have access efficiently to the video segment that corresponds to them.

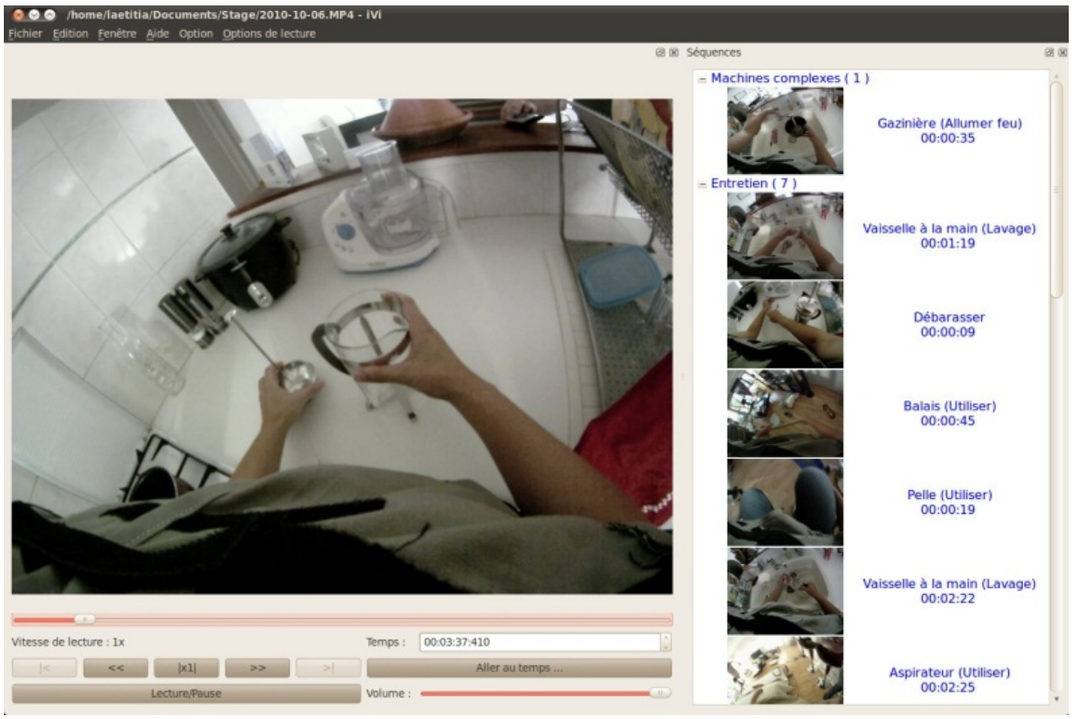


Figure 11: Consultation iVi interface

5.2.2.1 Functionalities

The software is composed of three parts (Figure 11). The video is presented on the left while the indexed content is shown on the right as temporal interval annotations. In this view, the activities are organized by category (Cleaning, Food, Hygiene, Leisure, etc.) and by decreasing order of activities length. Each activity is also represented by a snapshot which corresponds to the temporal center of this activity. The user is allowed to modify the temporal markers, the name of the activity or put an activity in another category. In the lower part the buttons to navigate in the video can be found. The users can navigate thanks to the slider, increase or decrease the video speed and the volume. Also, a specific time can be chosen in the video. Two buttons (|< and >|) allows to play the previous annotation or the next annotation of the current annotation.

5.2.2.2 Tool characteristics

The software is developed in C++ and use different libraries such as QT[4], VLC[5] and FFMPEG[6]. It's based on the iVi annotation software (presented in Section 4.2), by restricting and focusing the functionalities to those that are useful for an efficient visualisation.

Consultation iVi takes as input a XML file which contains the indexation of the activities, and output an activity report (Figures: 12, 13). This report contains all information about different activities of one video in the format (name, commentary, start, and end).

ACTIVITIES REPORT

Video studied : ZVNV_2

List : complex machines

Activity	Commentary	Start	End
Television (remote control)		00:02:13.016	00:03:34.231
Washing machine (Fill)		00:13:43.941	00:15:00.417
Oven (Light)		00:28:03.233	00:28:55.252
Gas cooker (Baking)		00:29:28.652	00:30:09.960

List : Cleaning

Activity	Commentary	Start	End
Broom(Use)		00:00:00.884	00:00:15.399
Hoover(Use)		00:16:00.744	00:20:04.421
Dishesbyhand(Washing)		00:36:43.220	00:37:02.306

Figure 12: Example of an activity report (.doc)

ACTIVITIES REPORT
Video studied : ZVNV_2
List : complex machines

[4] http://qt-project.org/doc/qt-4.8/

[5] http://tldp.org/REF/VLC-User-Guide/VLC-User-Guide.pdf

[6] http://ffmpeg.org/general.html

Activity ; Commentary ; Start ; End
Television (Remote control);;3986;6420;
Washing machine (Fill);;24693;26985;
Oven(Light);;50446;52005;
Gas cooker (Baking);;53006;54244;
List : Cleaning
Activity ; Commentary ; Start ; End
Broom (Use);;20023;23235;
Hoover (Use);;28793;36096;
Dishesbyhand (Washing);;66030;66602;

Figure 13: Example of an activity report (.csv)

5.2.2.3 Potential functionalities and Constraints

The *Consultation iVi* interface allows to visualize and to navigate in automatically indexed audio-visual content. This software is simple, and has a quick start. Thanks to the consultation iVi, we can observe and evaluate the results of ours algorithms of the activity recognition.

5.2.3 EyeTrackLab Visualisation Tool

UB1/LaBRI owns an eye tracker, that can be used to evaluate what part of images and videos are observed by a human observer. This device evaluates where the observer looks within the images, and represents this information as trajectories between fixations points, where the gaze temporarily makes a pause. This could be used to better define the saliency models for object detection and contextual information collection from visual data acquired within Dem@care, such as wearable and static video data. To process the eye tracker' data, we have developed a software EyeTrackLab to analyse and interpret these. This is an internal visualisation tool because it's dependent on this specific eye tracker.

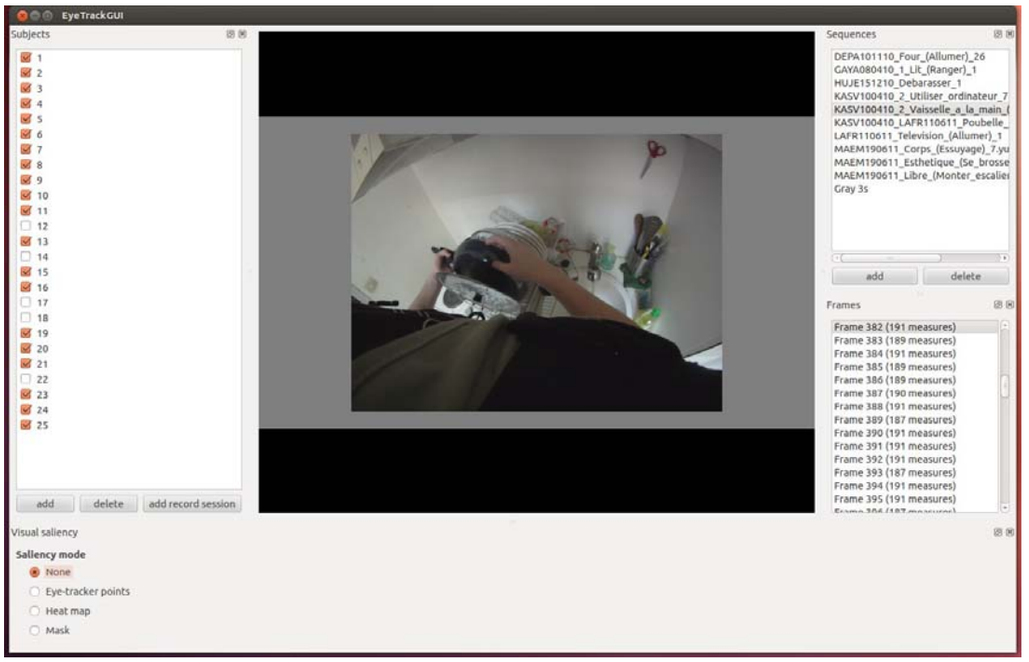


Figure 14: EyeTrackLab interface

5.2.3.1 Functionalities

The EyeTrackLab software allows visualizing the eye tracker data per person but also for a part or all persons. The interface is composed of five parts (see Figure 14). At the center, the image is presented. On the left, an interactive view allows the user to choose the subject who wishes to analyse data. The user can also choose to see data of more subjects. On the right, two views are available. The first represents the different video sequences. And the second represents all frames of one sequence selected. In this “Frames” view, the frame number and the measure number are shown. On the bottom, the display mode (saliency mode) can be selected, within eye-tracker points (discrete fixation points and trajectories), heat map (smoothed spatial distribution of the fixation points), or mask (alternative heat map visualisation), as illustrated in Figure 13.

5.2.3.2 Tool characteristics

The EyeTrackLab software is developed in C++ and depends on the QT library.

The software takes as input YUV videos and eye tracker’s data, and output as the saliency maps. An example of the output’ software is given in Figure 15. In this left figure, the red area represents the area which contains the more point. Also, we can say that people watch more in the object area rather than the whole image. People are attracted by objects and by the

hands. In the right figure, we can see a mask: the look is concentrated on the object (manipulating the dishes).

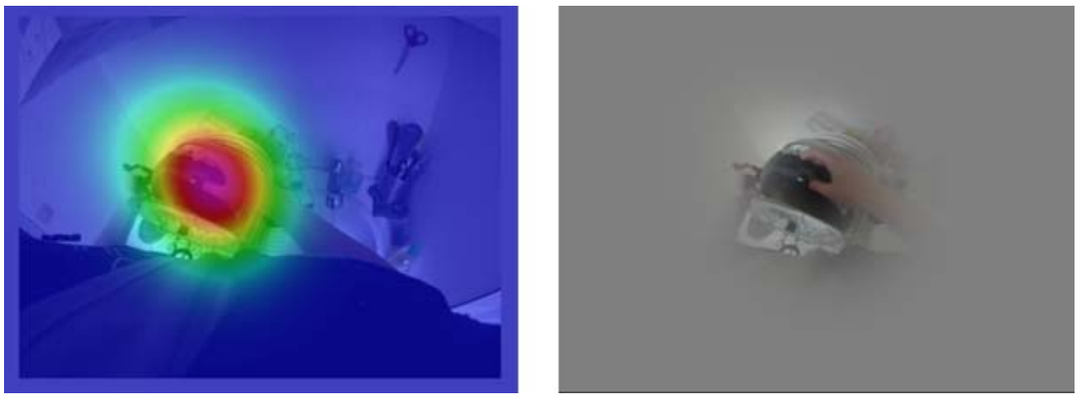


Figure 15: Saliency map and mask

5.2.3.3 Potential functionalities and Constraints

The EyeTrackLab visualisation tool is significant to analyze and interpret the eye tracker data. We can notice that the human vision is attracted by the motion and the object. So the saliency maps are related to the objects which are handled by the person during an activity. This tool could therefore support the development of new insights for the visual analysis of the video data acquired within Dem@Care (for the wearable camera data but also for fixed cameras).

6 Training Data Set

In this section, we first present the objectives and challenges of training process through the development of Dem@Care sensing components. Then specific needs (e.g. quantity - 20 samples for each type to be detected - and quality - sample resolution, precision of the annotation -) of annotated data required for training and performance evaluation process of Dem@care sensing components are described for each sensor.

6.1 Objectives and Challenges

The objective is to collect data for the extraction of relevant features of person's clinical status using a multi-sensor perspective. For this goal, audio and video recordings should be annotated by clinical experts and technical partners. The annotation needs to be in respect to relevant characteristics of person's clinical status (e.g., IADLs recognition in order to extract specific features in the realization of IADLs, speech features characteristic to the person's mood) with sufficient details (e.g., postures adopted, IADLs well/bad performed).). These can be used for training, for the evaluation of algorithms (Computer Vision and Machine Learning techniques) developed for the recognition of the annotated events and the classification of person's clinical status according the features extracted from audio/video fragment of interest. Training data set could be also useful for the validation of medical researches assumptions.

6.2 Needs

For multimodal capture of the same scene, the different sensor readings will be synchronized to allow for cross-modal analysis and annotation sharing. The need is to have the information necessary to put each video frame or audio segment onto the same timeline. Since most sensors have accurate quartz based clocks that do not drift significantly over the span of a few hours, an affine correction should be enough. For a given sensor, it could be defined as the frequency of acquisition of the sensor, and a time offset to align its own timeline to the reference timeline.

6.2.1 Video data

6.2.1.1 People detection and tracking from ambient video camera

Annotation type description: This task refers to the identification and the tracking over the time of physical object classified as “Person(s)” in the scene. Accurate detection of the person (i.e., bounding box perfectly enclosing the person’s body) is necessary for the correct estimation of person’s attributes (e.g., height, width, localization) over time, and may also impact in the person’s activity analysis.

The annotation of people detected/tracked is made at the frame resolution using a bounding box that encloses the person detected and tracked in the scene. A specific ID label is assigned to the person tracked. This bounding box is updated for each frame or interpolated over a time period (according the level of accuracy acceptable) in which the persons remains visible in the scene.

Objective of people detection/tracking annotations is to allow the evaluation of (1) the gait parameters assessment from participant’s global trajectory from video (can be used as ground truth), (2) the assessment of the detection/tracking vision modules, and (3) the assessment of the possible effects of People misdetection and tracking errors over automatic event detection.

Annotation needs: Initially the level of accuracy required for People detection/tracking would be the definition of bounding box enclosing the person,It should be noted that the Kinect RGBD data (depth map data) could be used as People detection and tracking ground truth data in absence of manually annotated data due to its tracking robustness to illumination changes.

6.2.1.2 Object detection from wearable video camera

Annotation type description: The purpose of object detection is to detect the presence of objects of daily living, such as a coffee machine, a book, and so forth, which are used by the person. Such information is of primary interest for the analysis of actions and activities, and provides input to the activity analysis system.

The objective of object annotation is to first construct the typology of the objects related to IADLs that are visible from the mobile camera, and second to evaluate the performance of the automatic object detection algorithms at detecting these instances.

Object annotation is described by a bounding box that encloses the object of interest, with a label stating its nature. This bounding box is updated for each frame in which the object remains visible.

Annotation needs: Within the data collection in controlled lab environment (described in section 7.2), objects of interest include objects that the person will interact with, or manipulate: table, phone, locker, kettle, television, paper, pillbox, basket and pitcher.
Because of the variability of appearance of objects during IADLs, annotation will target to label all videos of the CHUN scenario, in order to capture this variability in their natural context. When resolution permits, object annotation will be done on fixed videos as well.

6.2.1.3 Events recognition

Annotation type description: Events to detect are human actions and activities, essentially focused on the detection of Instrumental Activities of Daily Livings (IADLs) in order to assess the degree of autonomy of the older person.
The objective of event annotation is to first define an ontology of the events to recognize (e.g., person's state (posture/location), human action/activity) in the CHUN scenario that are relevant from clinical point of view, and second to model events using a generic ontology language (using a priori knowledge) or to build discriminative codebook for action/activity recognition (e.g., using classifier system based on trajectory features extracted from video sequences), and third to evaluate the performance of the automatic event recognition algorithm (uni-modal with only a sensor data type/or multi-modal approach) at detecting and the events instance.
Event annotation is described by a video segment tagging of the video sequence (time period delimitation: start/end time point) when the event occurred. Each event annotated is labeled with the corresponding name.

Annotation needs: Due to the difficulty to define the start/end time of human action/activity, different granularity levels describing the activity may be required. For the detection of IADLs within CHUN scenario, following events annotation may be required: (i) person's location in a zone of interest (e.g., inside Coffee zone, inside TV zone), (ii) interaction (different levels: touching, holding, using) with objects of interest needs to perform IADLs, and (iii) person's posture (i.e., standing, sitting, walking, bending, unknown). For the walking assessment within CHUN scenario, following events annotation may be required: (i) person's

location in zone of interest (i.e., “Walking zone”), (ii) person’s posture (i.e., standing, walking, doing a U-turn, unknown).

The annotation of events is not specific to fixed cameras, but can be shared with mobile cameras after synchronization (and other sensors), which provides a complementary point of view, and allows to refine these annotations. In particular, for interactions with objects, this point of view should be used to extend the annotation to cases where the fixed cameras may not have a good visibility (due to occlusion or poor resolution) on the IADLs of interest.

A high time resolution is required (i.e. more than 8 frames per second) due to the short duration of activities to perform during the short time frame allocated for each clinical scenario.

For action/activity recognition algorithm based on classification system of trajectory-based video features, specific requirements in terms of event annotation are specified:

- For actions/activities classification: A minimum of ten samples for each action/activity category is required for action/activity recognition module.
- For clinical status classification: A minimum of twenty samples by clinical status is required to be able to classify participant by clinical status from local features extracted from the video sequence of action/activity human. Also, to ensure a robustness of activity/action model to individual variations, the dataset of each clinical status should contain people of different shapes, sizes, genders and ethnicities.

6.2.2 Accelerometers

The annotation needs for accelerometer will be provided at event level. Video-based annotation will be translated to accelerometer data annotation based on their respective timestamps. More details about accelerometers measurements for specific purposes (e.g., gait analysis) could be also provided according to the project needs).

6.2.3 Audio sensor

Two levels of audio manual annotation have to be provided:

- *Task level annotation* characterizing relevant aspects of the task execution and associated with the respective audio file considered as a whole. For example: the number of errors in

counting aloud task or speech intelligibility level during the task performance. The task level annotation must be done immediately after the task execution.

- *Audio segment tagging* associated with certain segments inside the audio record and referring to the respective time-stamps. For example: speech passages exhibiting depressive state of mood of the participant within the recording of the discussion with the clinician.

6.2.3.1 Task level annotation

6.2.3.1.1 Speech rate

Annotation type description: Strictly speaking, speech rate is measured in number of words uttered per second on average. However for our purposes it is enough to estimate the speech rate subjectively.

Annotation needs: Following scale will be used: 1 – slow; 2 – normal; 3 – fast. Speech rate estimate will be done and annotated at the activity level: 1) once for the Directed Activities; 2) once for the Discussion with the Clinician (see Dem@Care deliverable D2.2 for clinical protocol description).

6.2.3.1.2 Speech intelligibility

Annotation type description: Speech intelligibility measures how clear (understandable) the participant speech is at the word level. In general the intelligibility it is not related to the level of comprehensiveness of the message conveyed. The message may be perfectly comprehensive (would it be available in the text form) while it is difficult to understand some words. On the opposite the words may be uttered very clear but the whole message does not make sense.

Annotation needs: Speech intelligibility will be estimated by the following scale: 1 – very low, most of the words are uttered unclear; 2 – low, many words are uttered unclear, understanding requires major effort; 3 – fair, some words are unclear, some effort required to understand them; 4- good, a few words are uttered not clear enough but still understandable; 5 – perfect, no effort is required to understand each and every word. Speech intelligibility estimate will be done and annotated at the activity level: 1) once for the Directed Activities; 2) once for the Discussion with the Clinician.

6.2.3.1.3 Correctness of vocal tasks

Annotation type description: This annotation type concerns with the measurement of correctness/adequacy of the execution of some tasks in the clinical assessment (e.g., correctness of the counting).

Annotation needs: The correctness annotation will be done for each task separately. Following correctness rate scale will be used for most of the tasks: 1 – very poor; 2 – poor; 3 – fair; 4 – good; 5 – very good/perfect. The correctness will be annotated for the following tasks as specified: Counting aloud (the number of errors in counting); Sentence repeating (approximate number of word errors); Articulation control (diadochokinetic test) (overall accuracy of the tokens uttering on the above scale); Verbal description of pictures (the description adequacy at the above scale).

6.2.3.1.4 Subjective cognitive load

Annotation type description: The aim is to rate the effort invested by the participant during the task execution as manifested in participant's speech, non-verbal sounds, facial expression, gestures, postures, etc. This effort is not necessarily related to the correctness of the execution. For example, it can be imagine, at least hypothetically, that a participant has performed the task without any error but it took much effort while another participant performed the task easily without any visible effort but absolutely incorrect.

Annotation needs: The subjective cognitive load will be measured on the following scale: 1 – no load observed, the patient did not exhibit any visible effort; 2 – minor load, soft rare indications; 3 – significant load, clear indications that the task posed difficulties; 4 – high load, consistent indications during the whole task execution; 5 – unaffordable load, the participant could not cope with difficulties, gave up. Subjective cognitive load will be estimated and annotated at the activity level: 1) once for the Directed Activities; 2) once for the Discussion with the Clinician.

6.2.3.1.5 Arousal/Interest and valence

Annotation type description: This type of annotation aims to support apathetic mood detection. The judgment will be based on participant's speech, facial expression, gestures, etc.

Annotation needs: The arousal level will be rated by the following scale: 0 – the participant is absolutely indifferent; 1 – low arousal, mostly indifferent but "wakes up" sometimes; 2 –

normal arousal, the level that would be expected from an average person passing an IQ test or something similar; 3 – high arousal, the participant is excited.

In general, the arousal level is independent of whether subject's attitude is either positive or negative (the valence). Although we do not expect the participants to exhibit the negative attitude, the arousal can be rated regardless of the valence. The valence will be rated independently by the following scale: -1 – negative; 0 – neutral; 1 – positive.

This type of annotation will be performed for: 1) Sentence repeating task; 2) Discussion with the Clinician activity.

6.2.3.2 Audio segment tagging

Annotation type description: This type of annotation requires listening to the audio recording, identifying in it certain events and marking the respective audio segments. The table below specifies the target events that will be identified and marked:

Table 27: Audio segment tagging description.

Non-verbal vocalizations
Laughter
Sighs
Other non-verbal vocalization
Salient exhibition of mood/emotion
Apathy/depression/sadness
Stress
Irritation
Other non-neutral emotional state
Salient indications of speech disfluency
Multiple false starts, multiple interjections, unintelligible passages

Annotation needs: Audio segment tagging will be applied to the recordings made during the Discussion with the Clinician. The number of audio segment tagged provided will be subjected to the CHUN capacity in terms of the workload.

6.2.4 Physiological data

No specific annotation is necessary for this data type.

7 Data Collection

This section describes benchmarking dataset available for each sensor (or multi-sensor platform), and its potential interests for Dem@Care project. Then tentative plans for data collection (for each sensor) during the pilots of the Dem@Care project are presented.

7.1 Benchmark data set

7.1.1 Video Monitoring Platform data set

Table 28: Benchmark datasets description

	Algorithms tested	**Relevance for Dem@Care**	**Comments**
Gerhome	• People detection, tracking and video event recognition. • Sensor stream filtering and contact event recognition. • Multimodal event recognition. • Multi-sensors analysis for everyday elderly activity monitoring.	High	• One person in the scene. • Multi-sensors (ambient and contact sensors).
Sweet-Home	• People detection/ tracking, action recognition and video Events recognition (events are: Activities of Daily living and physical activities). • Multimodal (video and actigraphy with posture detection data) event recognition.	High	• Scenario 01: Two persons (clinician and participant) inside the room/ Scenario 02 & 03: One person (participant) in the scene. • 2D video camera position: Occlusion problems making the IADLs recognition

			not feasible.
ADL dataset	• Events (events are: Activities of Daily Living) recognition.	High	• 2D video sensor only • One person in the scene
KTL dataset	• Action recognition.	Medium	• 2D video sensor only. • One person in the scene.
IMMED	• Activity recognition. • Location recognition. • Object detection. • Temporal segmentation.	High	• One person with a wearable camera, with one assistant.
Kit Robo-Kitchen	• Activities, objects recognition and detection.	Medium	-
TUM Kitchen	• Activities, objects recognition and detection.	Medium	-

Data set from GERHOME project[7]

Description about the experiments conducted with older persons can be found in [14].

- Description/Objective: The objectives of GERHOME project is to develop test and certify technical solutions for activity monitoring of the elderly at home, by using network of sensors.
- Location: GERHOME laboratory is a typical apartment of an elderly person: 41m² with an entrance, “a living room”, “a bedroom”, “a bathroom” and “a kitchen”.
- Population: 14 Older volunteers (no specific clinical characteristics in terms of cognitive disorders).

[7] http://gerhome.cstb.fr/

- Sensors: Video monitoring platform fed by a network of cameras and 24 environmental sensors (contact sensors, pressure sensors, water flow sensors, electrical sensors, presence sensors, Figures: 16-18.

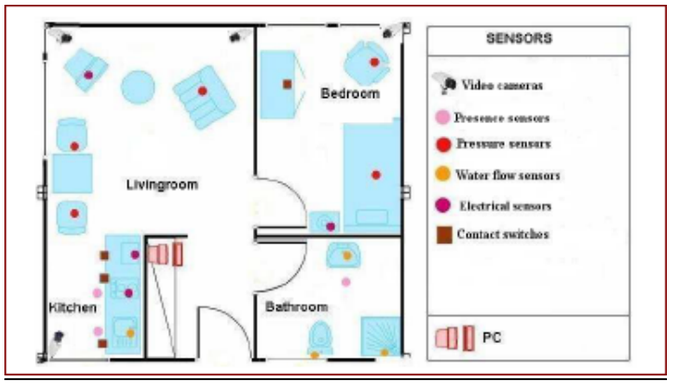


Figure 16: Overview of the instrumented Gerhome Laboratory

Figure 17: Views of Environmental sensors installed in the Gerhome laboratory.

In Figure 17: images from left to right: (a) Contact sensor on cupboard door in the kitchen , (b) Electrical sensor on electrical outlet in the kitchen, (c) Presence sensor in front of the washbowl in the bathroom, (d) Water sensor on water pipe in the kitchen, (e) Pressure sensor under the armchair in the living room.

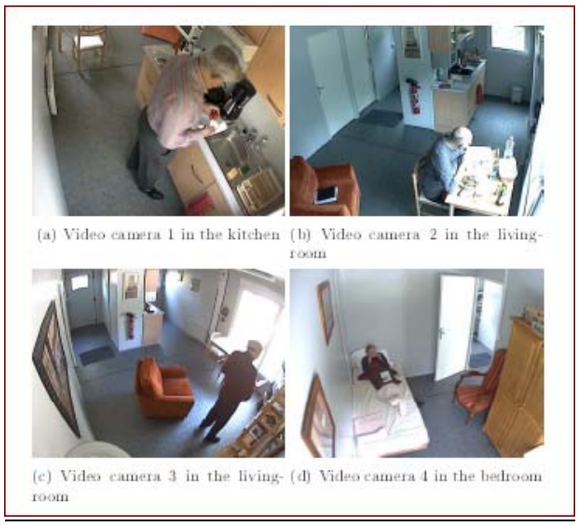


Figure 18: Views of the 4 video cameras inside the Gerhome laboratory.

- Scenarios: Older volunteers were alone in the apartment. Volunteers were encouraged to behave freely and to maintain as normal as possible their behaviours and were asked to perform a set of household activities such as “Preparing meal”, “taking meal”, “cleaning the kitchen”, “watching TV”, “taking a nap” More details about the activities description can be found in [14].
- Description of data set available: Each volunteer was observed during 4hours: 56 video sequences were acquired by 4 video cameras (~10frame per second), each video sequence contains 144 000 frames; and data from 24 environmental sensors were used. (24 environmental sensors). Activities of interest are body postures (e.g., “Standing”, ”Standing with Arms Up”, “Bending”, “Slumping”, “Sitting”…), kitchen activities (e.g., “Using Fridge”, “Using Stove”, “Preparing breakfast”, “Taking Meal”…); hygiene activities (e.g., “Washing hands or face”, ”Bathing”…), leisure activities (e.g., “Watching TV”, “Using Phone”, ”Reading paper/book/magazine”,…), bedroom activities (“Taking medication”, ”Sleeping”, ”Waking up”, ”Napping”) and other Activities (“Entering the house”, “Leaving the house”), [14].
- Images from the project: can be seen in APPENDIX A3.

Data set from SWEET-HOME project[8]

- Description/Objective: Propose a technological approach for behavioural assessment in Alzheimer's disease patients at early to moderate stages using video/audio/actigraphy sensing components.
- Location: Smart-room located in Memory Center of Nice (CHUN). The room of 32m² is equipped of household appliances need for the realization of Instrumental Activities of Daily Living (IADL). Assessment inside the smart room is proposed following a medical consultation.
- Population: Participants are aged more than 65 years with a Mini Mental Score Exam (MMSE) above 15. Ambulatory participants are classified in three clinical categories: Healthy Control participants / Patients with Mild Cognitive Impairment / Alzheimer's disease patients.
- Sensors: Video monitoring system is composed of 2D video camera ambient (8frames per second, 640×480 pixel resolution), and 3D video camera Kinect RGBD (~10acquisition per second). Audio is recorded with ambient micro. Wearable devices – Actigraphy Date: MotionPod (worn in the chest, 1data/second) and MotionLogger (worn on the wrist, 4data/minute).
- Scenarios: The clinical protocol takes place in the same scene (smart-home). Clinical protocol (2 Protocols *P1* and *P2* were conducted) followed by the participants is divided into three parts:

  - Scenario 01 ("Directed Activities", video sequence duration: ~5min): participants performed a battery of physical tests. Each physical battery test is composed of human actions (APPENDIX A4**,** Table A4.1). Each test is done at least one once (some participants don't manage to achieve the test). This scenario is intended to assess kinematic parameters about the participant's gait profile.

  - Scenario 02 ("Semi-directed Activities", video sequence duration: max. 20min (P1)/max. 15min (P2)) aims at evaluating the degree of independence of the participants. For that participants have to perform a list of Instrumental Activities of Daily Livings

[8] http://cmrr-nice.fr/sweethome/

(IADLs) with specific constraints for organizing them. All IADLs are not necessary performed by the participant (e.g. reasons: IADL's omission due to memory disorders, lack of time to perform the scenario) . Description can be found in APPENDIX A4**,** Tables: A4.2, A4.3.

- Scenario 03 ("Free activities", video sequence duration: 30min (P1)/5min(P2)) aims at assessing how the participant spontaneously initiates activities in the room (*e.g.*, reading magazines/newspapers, drinking, playing cards, and watching television) and organize his/her time without receiving specific instructions. There is no instruction: so all possible actions/activities which can be done in the room can be done by the participant many times.

- Description of data set available: For *P1*. 37 video sequences were recorded for each Scenario (01/02/03). For the 37 video sequences: time period delimitation of each "Physical battery test" (in the Scenario 01) and "IADLs" (when they occur in the Scenario 02) were done. No people tracking annotation is available.

  For *P2*: 82 recording sessions took place. 80 video sequences for the Scenario 01 (22 video sequences with the time period delimitation of each "Physical battery test" are available), 70 video sequences for the Scenario 02 (40 video sequences with time period delimitation of each "IADLs" are available), 80 video sequences for the Scenario 03 (no event annotation is available). No people tracking annotation is available.

  The SEET-HOME corpus is annotated in terms of activities composing the "Physical battery test" and the IADLs performed within the Scenario 02 (see APPENDIX A4, Tables: A4.1, A4.2, A4.3).
- Images from the project: can be seen in APPENDIX A5.

ADL data set[9]

- Description/Objective: The University of Rochester Activities of Daily Living dataset (ADL dataset) contains ten types of human activities of daily living that people have to perform three times. These activities were selected to be useful for an assisted cognition task and to be difficult separate on the basis of any single source of information.

[9] http://www.cs.rochester.edu/~rmessing/uradl/

- Location: Videos were recorded inside a room in front of the location where activities were realized. Videos were recorded from about two meters away by the ambient 2D video camera.
- Population: Five people of different shapes, sizes, gender and ethnicities.
- Sensors: Ambient 2D video camera (30 frames per second, 1280 x 720 pixel resolution).
- Scenarios: The full list of activities is: answering a phone, dialing a phone, looking up a phone number in a telephone directory, writing a number on a whiteboard, drinking a glass of water, eating snack chips, peeling banana, eating a banana, chopping a banana, and eating food with silverware. Each activity is performed in the same scen (kitchen room) three times by people of different shapes, sizes, genders, and ethnicities.
- Description of dataset available: 150 video sequences
- Images from the project: can be seen in APPENDIX A6.

KTL data set[10]

- Description/Objective: KTL dataset contains six types of human actions performed several times by different subjects and for different scenarios.
- Location: Four contextual scenes were tested: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), indoors (s4).
- Population: 25 people of different shapes, sizes, gender.
- Sensors: Ambient 2D video camera (25 frames per second, 160 x 120 pixel resolution)
- Scenarios: The full list of human actions is: walking, jogging, running, boxing, hand waving, and hand clapping. Each action is performed several times by 25 persons in the four contextual scenes: (s1), (s2), (s3) and (s4).
- Description of dataset available: 2391 sequences organized in 599 video files (duration of sequences is 4seconds in average)
- Images from the project: can be seen in APPENDIX A7.

[10] http://www.nada.kth.se/cvap/actions

Data set from IMMED project.

- Description/Objective: The IMMED project aims at developing new tools for indexing activities of daily living in videos obtained from wearable cameras. In the context of dementia diagnosis by doctors, patient activities are recorded in the environment of their home using a lightweight wearable device, to be later visualized by the medical practitioners. This recording mode provides challenging video data, consisting in a single sequence shot where strong motion and sharp lighting changes often appear. Because of the length of the recordings, tools for an efficient navigation in terms of activities of interest are crucial.
- Location: Person home.
- Population: Healthy Control participants / Patients with Mild Cognitive Impairment
- Sensors: Wearable camera worn on the shoulder that acquires video and audio.
- Scenarios: The acquisition is supervised by an assistant, during a planned session of up to 1 hour. The scenario is divided in two parts: first the assistant ask the person to present all rooms of the apartment, second the assistant proposes directed activities to be accomplished by the person. The first part is rather short (less than 5 minutes) and constitutes a bootstrap video to be annotated that is used to model the environment. The secondis longer (between 30 minutes and 1 hour), and is the data to be analyzed with respect to activities of daily living. The scenario is not strictly defined, as the assistant has to select and adapt the proposed activities to each specific home environment and to the capacities of the person during the acquisition session.
- Description of data set available: This corpus currently contains 53 videos recorded by 44 patients and 7 healthy volunteers for a total of 21h08.

  Note: This unique dataset is at the moment restricted to UB1 and CHU Bordeaux access, due to protocols validated by the ethical committee (CPPRB) and data privacy committee (CNIL) to which the exploitation of such patient data was declared in the name of the partners of the IMMED project. It is planned that the official promoter of the study (CHU Bordeaux) will file a request on next September to obtain the permission to share this data within Dem@care.

The IMMED corpus is annotated in terms of objects, localization and activities, which are now detailed:

For the objects, eleven objects of interest are considered: coffee maker, sink, television, basin, gas-cooker, phone, pillbox, microwave, oven, washing machine, electric plate. A description is given in Table 29:

Table 29: Number of frames per object

	Coffee maker	**Sink**	**TV**	**Basin**	**Gas-cooker**	**Phone**	**Pillbox**	**Micro-wave**	**Oven**	**Washing machine**	**Electric plate**	**TOTAL**
Frames no.	240	659	603	622	373	387	75	315	106	75	120	3575

Concerning the annotation of the location, six types of places are defined: bedroom, bathroom, kitchen, living-room, other, outside. When several rooms correspond to the same type of place, they receive the same annotation.

The activities of interest correspond to the Instrumental Activities of Daily Leaving (IADL) executed by a patient (or a healthy volunteer) according to the taxonomy and scenario defined by a medical practitioner. The set of activities is «Food Manual Preparation», «Food Drink», «Displacement Free», «Cleaning Hoover», «Cleaning Broom», «Cleaning Clear», «Cleaning Bed», «Cleaning Shovel», «Cleaning Dustbin», «Cleaning Dishes by hand», «Hygiene Body», «Hygiene Beauty», «Hygiene Clothes», «Leisure Gardening», «Leisure Reading», «Leisure Computer», «Complex machines CoffeeMaker», «Complex machines GasCooker», «Complex machines WashingMachine», «Complex machines Microwave», «Complex machines Television», «Medicines Medicines», «Relationship Phone», «Budget Budget»…

- Images from the project: can be seen in APPENDIX A8.

The KIT Robo-Kitchen Activity Data Set[11]

- Description/Objective: The KIT Robo-Kitchen data set is designed for the application in Human Robot Interaction (HRI) scenarios.
- Description of data set available: The Robo-kitchen data set consists of 14 typical kitchen activities recorded in two different stereo-camera setups, and each performed by 17 subjects (APPENDIX A9). It is composed of 37 videos which each represent an activity such as «Cut», «Dry», «Peel», «Wash», «Coffee».

  Each video correspond to an activity, for which we have annotated the interest objects. Eleven objects are retained: boiler, bowl, dishwasher, frying-pan, gas-cooker, microwave, mug, oven, plate, sink, and table. More details are given in Table 30. For this corpus all frames have been annotated.

Table 30: Frames number per object for Robo-kitchen.

	Boiler	**Bowl**	**Dishwasher**	**Frying-pan**	**Gas-cooker**	**Microwave**	**Mug**	**Oven**	**Plate**	**Sink**	**Table**	**TOTAL**
Frames no	2730	1590	825	5160	18329	14398	3330	14398	25242	12314	14444	112 760

- Images from the project: can be seen in APPENDIX A9.

TUM Kitchen Data Set[12]

- Description/Objective: The TUM Kitchen Data Set is provided to foster research in the areas of markerless human motion capture, motion segmentation and human activity recognition. The recorded activities have been selected with the intention to provide realistic and seemingly natural motions, and consist of everyday manipulation activities in a natural kitchen environment.
- Images from the project: can be seen in APPENDIX A10.

[11] http://cvhci.anthropomatik.kit.edu/projects/act/kitchen/

[12] http://ias.in.tum.de/software/kitchen-activity-data

7.2 First Dem@Care Data Collection: in a controlled lab environment.

Data collection activities in a controlled lab environment will be performed in CHU Nice, as a first step towards Pilot 1 for Dem@Care project, as described in D1.2: 1st Interim Management Report, in section 3, WP8/ Task 8.3: Pilot for Assisted Living in France.

In the first data collection activity planed for the summer of 2012, 150 participants aged more than 65years are expected to be recruited, with the following repartition by clinical status: Healthy Control participants (n=50), patients with Alzheimer's disease at pre-dementia stage (n=50), patients with Alzheimer's disease at dementia stage (n=50). Age and gender matching between clinical status groups will be aimed. The consultation will be divided into three parts: (T1) Consultation with the physician, (T2) Clinician consultation with a neuropsychologist, (T3) Ecological Assessment in the experimental room with the sensors recording. All data collected during the consultation of participant into Dem@Care. Important components of this data collection in controlled lab environment in Nice are described below.

7.2.1 Participant's profile

During the phases T1 and T2, the following demographical and clinical will be collected:

Table 31: Participants characteristics

Demographical characteristics	**Clinical characteristics**
• Gender • Date of birth • Education Level • Laterality • Size	• Diagnosis established the day of the recording session. Participant is assigned to one of this 4 categories: Healthy Control participants, Alzheimer's disease at pre-dementia stage, Alzheimer 's disease at dementia stage, Other dementias [1,9]. • Cognitive abilities assessment - Mini-Mental State Exam (MMSE) [6]. - Frontal Assessment Battery (FAB) [3]. - Trail making test A and B [10]. - Short Cognitive Battery [12]. - The Free and Cued Selective Reminding Test [7].

	• Neuropsychiatric/Mood assessments - NPI [2] - DSM-IV Criteria for depression - Apathy Inventory (AI) [11] and diagnostic criteria for apathy [13]. • Motricity abilities assessment - Part III of the Unified Parkinson's Disease Rating Scale (UPDRS) [5]. • Autonomy assessment - Instrumental Activities of Daily Living for Elderly (IADL-E) [8].

7.2.2 Sensors data

A description of the data set to be collected by sensor devices is provided. Training data with object/person detection/event recognition will be provided by clinical and technical partners (audio and video sensor). Computers used for data acquisition will be synchronized using Network Time Protocol (NTP) to allow further synchronization between all sensor data.

7.2.2.1 Ambient video camera

Table 32: Description of the data set to be collected by ambient video camera.

Sensors	2D video camera	Kinect RGBD
Number of devices Used/recording session	2	2
Sensor position	Inside the room to capture activities undertaken by the participant	Inside the room to capture activities undertaken by the participant
Number Recording sessions expected (number/clinical status)	150 (50 by each clinical status category)	150 (50 by each clinical status category)

7.2.2.2 Wearable video camera

Table 33: Description of the data set to be collected by wearable video camera.

Sensors	Wearable video camera
Number devices Used/recording session	1
Sensor position	Worn on the shoulder
Number of Recording sessions expected (number/clinical status)	150 (50 by each clinical status category)

7.2.2.3 Wearable still image camera (SenseCam)

Table 34: Description of the data set to be collected by SenseCam.

Sensors	Wearable still image camera (SenseCam)
Number devices Used/recording session	1
Sensor position	Worn on lanyard around the neck
Number of Recording sessions expected (number/clinical status)	Not decided: To be defined Not a compulsory device.

7.2.3 Audio data collection

Table 35: Description of the data set to be collected by audio sensor device.

Sensors	Wearable microphone
Number devices Used/recording session	2 (1 for clinician+1 for participant)
Sensor position	Next to the mooth
Number Recording sessions expected (number/clinical status).	150 (50 by each clinical status) * At this date: no material was received, therefore the number expected will be lower.

An ambient microphone will be also used during the protocol.

7.2.4 Accelerometer data collection

Table 36: Description of the data set to be collected by accelerometers.

Sensors	WIMU devices
Number devices Used/recording session	Not decided: to be defined.
Sensor position	Not decided : to be defined
Number Recording sessions expected (number/clinical status)	Not decided: To be defined * At this date: the WIMU devices are still in evaluation phase. Therefore the number of recording that would be available at the end this data collection activity in Nice, cannot be estimated.

7.2.5 Physiological data collection

Table 37: Description of the physiological data set.

Sensors	BodyMedia SenseWear Pro3	Philips DTI-2
Number devices Used/recording session	1	1
Sensor position	Upper the arm	On the wrist
Number Recording sessions expected (number/clinical status)	Not decided: to be defined Not a compulsory device.	Not decided: to be defined Not a compulsory device. * At this date: no material was received, therefore the number expected will be lower.

7.3 Commentary about other Dem@Care Pilots

At this time period of Dem@Care project (Month 8), the scenarios conducted in Home and Nursing Home pilots are not formalized but the sensors described in this report should be also compatible with them. The results of the data collection activity in Nice , which is part of Pilot 1, are going to be used as reference to evaluate what can and cannot be feasibly

measured in uncontrolled environment besides to the activities that are only performed by participants in uncontrolled environment.

8 Conclusions

This deliverable provides information about the data collection (from the acquisition to the processing by manual annotation used for sensing component evaluation) in order to define the needs for the development of a multi-sensing system enabling the assessment of behavior and cognitive disorders in persons suffering from Alzheimer's Disease at pre-dementia and dementia stages during the Lab-based Pilot.

This deliverable details the sensors specification used to capture participant's activities during the data collection activities in Nice which is part of Pilot 1 of the project. To assure the multi-sensor approach, we highlight that at each sensor reading a timestamp value should be provided. This timestamps will be used as the basis for time synchronization. For a full activity representation, we suggest that sensors acquisition rate should be of at least 4 readings/sec.

Annotation and visualisation tools developed by Dem@Care project partners are described and their potential application for Dem@Care is explored. Benchmark data sets are also presented to support the analysis of the constraints that should be taken in account in the sensing components development, and for performance evaluation of Dem@Care sensor system in respect to existing systems. Finally, specific details about the process for data collection in controlled lab environment are presented.

The variety of tools also highlights the need of a common file format for sensor readings representation and data sharing.

9 References

[1] Albert M.S., et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement, 2011. 7(3): p.270-9

[2] Cummings, J.L., et al., The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. Neurology, 1997. 48 (suppl 6): s10-s16.

[3] Dubois, B., et al., The FAB: a Frontal Assessment Battery at bedside. Neurology, 2000. 55(11): p. 1621-6.

[4] Dubois B., et al., Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol, 2007. 6(8): p. 734-46.

[5] Fahn S, Elton RL, UPDRS program members. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden CD, Goldstein M, Calne DB, editors. Recent Developments in Parkinson's Disease, vol. 2. Florham Park, NJ: Macmillan Healthcare Information, 1987. p. 153–163, 293–304.

[6] Folstein, M.F., et al., "Mini-mental test". A practical method for grading the cognitive state of patients for the clinician. J Psychiatry Res, 1975. 12: p. 189-198.

[7] Grober, E., et al., Genuine memory deficits in dementia. Developmental Neuropsychology, 1987. 3: p. 13-36.

[8] Mathuranath P.S., et al., Instrumental activities of daily living scale for dementia screening in elderly people. Int Psychogeriatr, 2005. 17(3): p. 461-74.

[9] McKhann G.M., et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement, 2011. 7(3): p. 263-9. Epub 2011 Apr 21.

[10] Reitan R.M., The relation of the trail making test to organic brain damage. J Consult Psychol, 1955. 19(5): p. 393-4.

[11] Robert, P.H., et al., The apathy inventory: assessment of apathy and awareness in Alzheimer's disease, Parkinson's disease and mild cognitive impairment." Int J Geriatr Psychiatry, 2002. 17(12): p. 1099-105.

[12] Robert P.H., et al., Validation of the Short Cognitive Battery (B2C). Value in screening for Alzheimer's disease and depressive disorders in psychiatric practice. Encephale, 2003. 29(3 Pt 1): p. 266-72.

[13] Robert, P.H., et al., Proposed diagnostic criteria for apathy in Alzheimer's disease and other neuropsychiatric disorders. Eur Psychiatry, 2009. 24(2): p. 98-104.

[14] Zouba-Valentin N, “Multisensor Fusion for Monitoring Elderly Activities at Home", PhD thesis, Nice-Sophia Antipolis University, January 2010.

[15] ETHOWARCHER annotation tool**:** http://www.ethowatcher.ufsc.br/**.**

[16] Etholog annotation tool: http://www.ip.usp.br/docentes/ebottoni/EthoLog/ethohome.html.

[17] Observer XT annotation tool: http://www.noldus.com/human-behavior-research/products/the-observer-xt.

[18] Description of file formats for ViSEvAl visualisation software http://team.inria.fr/stars/2012/02/02/viseval-software/.Potential functionalities and Constraints.

[19] Official GoPro documentation: http://gopro.com/.

10 APPENDIX

A.1. APPENDIX A1

ViPER-GT Tool - XML file describing the different values annotated by the user.

```
<config>
	<descriptor name=”physical_object” type=”OBJECT”>
		<attribute dynamic=”false” name=”type” type=”http://...#lvalue”>
			<data:lvalue-possibles>
				<data:lvalue-enum value=”NULL”/>
				<data:lvalue-enum value=”Person”/>
				<data:lvalue-enum value=”Vehicle”/>
				<data:lvalue-enum value=”Equipment”/>
			</data:lvalue-possibles>
		</attribute>
		<attribute dynamic=”false” name=”subtype” type=”http://...#lvalue”>
			<data:lvalue-possibles>
				<data:lvalue-enum value=”NULL”/>
				<data:lvalue-enum value=”GROUND_POWER_UNIT”/>
				<data:lvalue-enum value=”PS”/>
				<data:lvalue-enum value=”LS”/>
				<data:lvalue-enum value=”PW”/>
				<data:lvalue-enum value=”PERSON”/>
			</data:lvalue-possibles>
		</attribute>
		<attribute dynamic=”true” name=”info2D” type=”http://...#bbox”/>
		<attribute dynamic=”true” name=”GP_AFT” type=”http://...#point”/>
		<attribute dynamic=”true” name=”GP_FWD” type=”http://...#point”/>
	</descriptor>

	<descriptor name=”event” type=”OBJECT”>
		<attribute dynamic=”false” name=”name” type=”http://...#lvalue”>
			<data:lvalue-possibles>
				<data:lvalue-enum value=”GPU_Positioning”/>
				<data:lvalue-enum value=”Handler_Deposites_Chocks”/>
				<data:lvalue-enum value=”Aircraft_Arrival”/>
				<data:lvalue-enum value=”PBB_Positioning”/>
				<data:lvalue-enum value=”AFT_CN_Loading_Operation”/>
				<data:lvalue-enum value=”AFT_CN_Unloading_Operation”/>
```

```
                    </data:lvalue-possibles>
                </attribute>
                <attribute dynamic=”false” name=”physical_object” type=”http://...# dvaluelist”/>
                <attribute dynamic=”false” name=” contextual _object” type=”http://...# dvaluelist”/>
        </descriptor>
        <descriptor name=”Information” type=”FILE”>
                <attribute dynamic=”false” name=”SOURCETYPE” type=”http://...#lvalue”>
                    <data:lvalue-possibles>
                        <data:lvalue-enum value=”SEQUENCE”/>
                        <data:lvalue-enum value=”FRAMES”/>
                    </data:lvalue-possibles>
                </attribute>
                <attribute dynamic=”false” name=”NUMFRAMES” type=”http://...#dvalue”/>
                <attribute dynamic=”false” name=”FRAMERATE” type=”http://...#fvalue”/>
                <attribute dynamic=”false” name=”H-FRAME-SIZE” type=”http://...#dvalue”/>
                <attribute dynamic=”false” name=”V-FRAME-SIZE” type=”http://...#dvalue”/>
                <attribute dynamic=”false” name=”FIRST-FRAME” type=”http://...#dvalue”/>
                <attribute dynamic=”false” name=”CAMERA” type=”http://...#lvalue”>
                    <data:lvalue-possibles>
                        <data:lvalue-enum value=”JAI1”/>
                        <data:lvalue-enum value=”JAI2”/>
                        <data:lvalue-enum value=”JAI3”/>
                    </data:lvalue-possibles>
                </attribute>
        </descriptor>
    </config>
<data>
        <sourcefile filename=”file:/D:/videos/Co-friend/MPEG2/COF-8/COF-8_JAI4.mpg”>
                <file id=”0” name=”Information”>
                    <attribute name=”SOURCETYPE”/>
                    <attribute name=”NUMFRAMES”>
                        <data:dvalue value=”86259”/>
                    </attribute>
                    <attribute name=”FRAMERATE”>
                        <data:fvalue value=”1.0”/>
                    </attribute>
                    <attribute name=”H-FRAME-SIZE”>
                        <data:dvalue value=”704”/>
                    </attribute>
                    <attribute name=”V-FRAME-SIZE”>
                        <data:dvalue value=”576”/>
```

```
                </attribute>
                <attribute name=”Camera”>
                        <data:lvalue value=”JAI4”/>
                </attribute>
        </file>

        <object framespan=”23100:23420” id=”0” name=”physical_object”>
                <attribute name=”type”>
                        <data:lvalue value=”Vehicle”/>
                </attribute>
                <attribute name=”subtype”>
                        <data:lvalue value=”TRANSPORTER”/>
                </attribute>
                <attribute name=”info2D”>
                        <data:bbox framespan=”23100:23101” height=”85”
                                width=”102” x=”64” y=”228”/>
                        <data:bbox framespan=”23102:23102” height=”86”
                                width=”103” x=”65” y=”228”/>
                        <data:bbox framespan=”23103:23103” height=”87”
                                width=”103” x=”66” y=”229”/>
                </attribute>
                <attribute name=”GP_AFT”>
                        <data:point framespan=”23100:23101” x=”83” y=”262”/>
                        <data:point framespan=”23102:23102” x=”84” y=”263”/>
                        <data:point framespan=”23103:23103” x=”85” y=”263”/>
                </attribute>
                <attribute name=”GP_FWD”>
                        <data:point framespan=”23100:23100” x=”131” y=”301”/>
                        <data:point framespan=”23101:23101” x=”132” y=”302”/>
                        <data:point framespan=”23102:23102” x=”133” y=”303”/>
                </attribute>
        </object>

        <object framespan=”23100:23420” id=”0” name=”event”>
                <attribute name=”name”>
                        <data:lvalue value=”GT_STD”/>
                </attribute>
                <attribute name=”physical_object”>
                        <data:dvaluelist list=”122,33,59,88”/>
                </attribute>
                <attribute name=”contextual_object”>
```

```
                                    <data:dvaluelist list="1,5,7,12"/>
                                </attribute>
                            </object>
                        </sourcefile>
</data>
```

A.2. APPENDIX A2

iVi Tool – Formats of the XML files used for storing annotations.

- The first format ("XML only") is composed of a XML file which contains the complete annotation information (see Figure A2.1). The activities annotations are between the markers <listsTemporelAnnotation> and </listsTemporelAnnotation>. And the objects annotations are between the markers <listsSpatialAnnotation> and </listsSpatialAnnotation>. Each temporal annotation contains the category name (<name>), the activity name (<annotation>), the start frame index (<start>), the end frame index (<end>) and a comment (option). Each spatial annotation contains the object name (<name>), the sequence name (<annotation>), the start frame index (<start>) , the end frame index (<end>) and the coordinates of the bounding box (<position>) : the coordinates (x,y) of the top left point and (w,h) the width and height of the bounding box.
- The second format ("XML+TXT") uses a XML file that contains the path of a txt file which contains the annotation information. There is one txt file per activity and per object (see Figure A2.2). The XML file only contains the category/object name <name> and the path of the txt file (<file>). The other information are saved in txt file following the following ASCII format :

  - A txt activity file contains for each activity segment:

```
StartFrame EndFrame ActivityName comment (optional)
```

- A txt object file contains for each bounding box:

```
StartFrame EndFrame x y width height  SequenceName comment (optional)
```

```
<!DOCTYPE project PUBLIC '-//XMLDTD//' 'projectXml.dtd'>
<!—This file content an video annotation file with Xml lists→
<project type="xml">
 <video>
  <file>/home/laetitia/Documents/mnt/PEPS/Data/Video/TournagesIMMED/VolontairesSains/ZVNV/ZVNV_2.MP4</file>
  <fps>29.97</fps>
  <videoSize>77580</videoSize>
 </video>
 <listsTemporelAnnotation>
  <list>
   <name>Complex Machines</name>
```

```
  <item>
   <annotation>Television (Remote control)</annotation>
   <start>3986</start>
   <end>6420</end>
  </item>
 <list>
  <name>Cleaning</name>
  <item>
   <annotation>Broom (Use)</annotation>
   <start>26</start>
   <end>461</end>
  </item>
  <item>
 <list>
  <name>Food</name>
  <item>
   <annotation>Manual preparation (Cut)</annotation>
   <start>58029</start>
   <end>61196</end>
  </item>
  <item>
</list>
…
</listsTemporelAnnotation>
 <listsSpatialAnnotation>
  <list>
  <name>tv</name>
  <item>
   <annotation>Séquence1</annotation>
   <start>4029</start>
   <end>4029</end>
   <position>379 5 430 244</position>
  </item>
  <item>
   <annotation>Séquence2</annotation>
   <start>4030</start>
   <end>4030</end>
   <position>420 0 351 240</position>
  </item>
 <list>
  <name>oven</name>
  <item>
   <annotation>Séquence1</annotation>
   <start>50542</start>
   <end>50546</end>
   <position>760 31 520 322</position>
  </item>
</list>
</listsSpatialAnnotation>
</project>
```

Figure A2.1- Example of a XML file in the “XML only format”

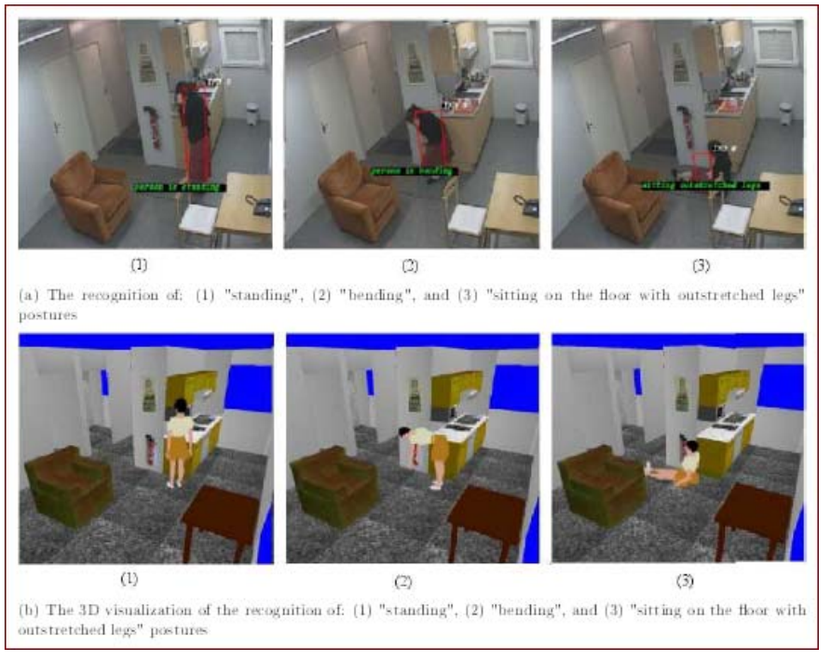
```
<!DOCTYPE project PUBLIC ‘-//XMLDTD//’ ‘projectTxt.dtd’>
<!—This file content an video annotation file with Txt lists→
<project type=”text”>
<video>
<file>/home/laetitia/Documents/mnt/PEPS/Data/Video/TournagesIMMED/VolontairesSains/ZVNV/ZVNV_2.MP4</file>
<fps>29.97</fps>
<videoSize>77580</videoSize>
</video>
<listsTemporelAnnotation>
<list>
<name>Complex Machines</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_complex
Machines.txt</file>
</list>
<list>
<name>Cleaning</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_Cleaning.txt</file>
</list>
```

```
<list>
<name>Foode</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_Food.txt</file>
</list>
<list>
<name>Hygiene</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_Hygiene.txt</file>
</list>
<list>
<name>Leisure</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_Leisure.txt</file>
</list>
</listsTemporelAnnotation>
<listsSpatialAnnotation>
<list>
<name>tv</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_tv.txt</file>
</list>
<list>
<name>oven</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_oven.txt</file>
</list>
<list>
<name>sink</name>
<file>/home/laetitia/Documents/Stage/IMMED/AnnotationsGT/TournagesIMMED2010/ZVNV/txt/ZVNV_2_sink.txt</file>
</list>
</listsSpatialAnnotation>
</project>
```

Figure A2.2- Example of XML file in the “XML+TXT” format

A.3. APPENDIX A3

Recognition and the 3D visualisation of some human activities recognised, GERHOME project (illustrations extracted from PhD thesis of Zouba N., 2010).



(a) The recognition of: (1) "standing", (2) "bending", and (3) "sitting on the floor with outstretched legs" postures

(b) The 3D visualization of the recognition of: (1) "standing", (2) "bending", and (3) "sitting on the floor with outstretched legs" postures

A.4. APPENDIX A4

Description of Human actions/activities performed by participants of Sweet-Home project

Physical battery tests	**Human actions composing the physical battery test**	**Occlusion problems**
Balance test	- Standing with feet side by side - Standing with feet in semi-tandem position - Standing with feet in tandem position - Standing on the right foot - Standing on the left foot	Low
Walking test	- Walking from the face - Walking from the back	Low
Transfer test	- Sitting down - Standing up	Low
Test Up and Go (TUG) test	- Standing up - Walking from the face - doing an U-turn - Walking from the back - Sitting down	Low

Table A4.1

IADLs	**Human actions and/or Object Interaction associated with IADLs**	**Occlusion problems**
Reading	- Holding newspaper - Looking in the newspaper	Low
Turning on tea kettle	- Tea kettle turning on	High (Person is back)
Calling somebody	- Dialing phone number - Holding the phone handset	Low
Watering the plant	- Holding watering can	High (Person is back)
Watching TV	- Holding the remote control	Medium

Classifying playing cards	- Holding playing cards	Low
Matching ABCD sheets of paper in the room	- Holding sheets of paper in the hand	Medium (for start/end time: the person is back)

Table A4.2

IADLs	**Human actions and/or Object Interaction associated with IADLs**	**Occlusion problems**
Reading	- Holding newspaper in the hand	Low
Watching TV	- Holding remote control - TV turning on/off (length of events 2frames)	Medium
Watering the plant	- Holding watering can - Adding water on the plant	High (person is back)
Preparing tea	- Tea kettle turning on - Holding tea kettle	High (person is back)
Calling somebody	- Dialing phone number - Holding the phone handset	Low
Answering the phone	- Holding the phone	Low
Preparing drug box	- Catching prescription/drugs can from the drug box - Putting away prescription/drugs can from the drug box - Selecting drugs - Putting drugs into the pillular	Low
Writing shopping list	Writing on the sheet	Medium (confusing to see on what the person writes)
Paying Electricity bill	Writing on the check	

Table A4.3

A.5. APPENDIX A5

Images from the Sweet-Home project

a) Scenario 01: “Walking from the back” during the “Walking test”

b) Scenario 01: “Walking from the front” during the “Walking test”

c) Scenario 01:” Walking from the front” during the TUG test

d) Scenario 01: “Doing an U-turn” during the TUG test

e) Scenario 01: “Sitting down” during the “Transfer test”

f) Scenario 02: “Matching ABCD sheets of paper in the room”

g) Scenario 02: “Reading”

h) Scenario 02: “Reading”

A.6. APPENDIX A6

Samples of ten ADLs (ADL dataset, downloaded from the official website).

a) ADL: Answering phone

b) ADL: Chopping banana

c) ADL: Dialling phone

d) ADL: Drinking water

e) ADL: Eating banana

f) ADL: Eating chips snack

g) ADL: Looking up in a phone directory

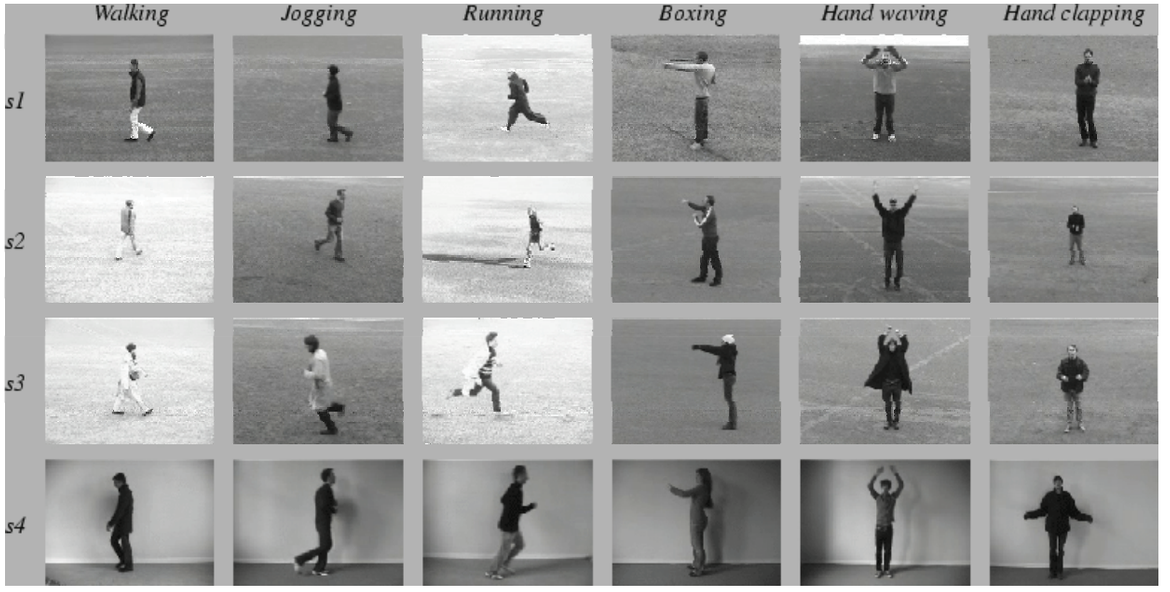
h) ADL: Peeling banana

i) ADL: Using silverware

j) ADL: Writing on whiteboard

A.7. APPENDIX A7

Samples of 6 types of human actions in 4 different contextual scenes (KTL dataset, downloaded from the official website).



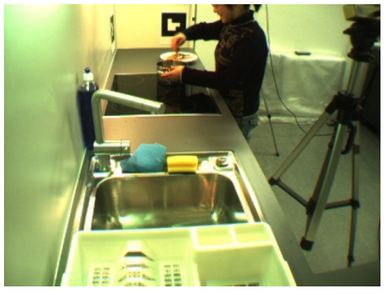
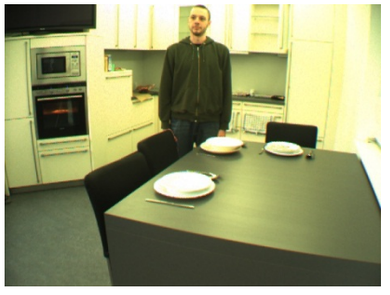
A.8. APPENDIX A8

Images from the IMMED corpus.



A.9. APPENDIX A9

Images from the Kit Robo-Kitchen corpus


A.10. APPENDIX A10

Image from the TUM Kitchen corpus