CIRDO: Smart companion for helping elderly to live at home for longer

Saïda Bouakaz, Michel Vacher, Marc-Eric Bobillier-Chaumon, Frédéric Aman, Salima Bekkadja, François Portet, Erwan Guillou, Solange Rossato, Elodie Desserée, Pierre Traineau, et al.

HAL Id: hal-00967911

https://hal.archives-ouvertes.fr/hal-00967911

Submitted on 2 Apr 2014

CIRDO : Smart Companion for helping elderly to live at home for longer ✩

S. Bouakaz[a,*], M. Vacher[b], M.-E. Bobillier Chaumon[c], F. Aman[b], S. Bekkadja[c], F. Portet[b], E. Guillou[a], S. Rossato[b], E. Desserée[a], P. Traineau[d], J.-P. Vimont[1], T. Chevalier[f]

[a] *LIRIS, UMR 5205 CNRS/Université Claude Bernard Lyon 1/ 43, boulevard du 11 novembre 69622 Villeurbanne, France*
[b] *LIG, UMR5217 CNRS/UJF/Grenoble-INP, 41 rue des Mathématiques, 38041 Grenoble, France*
[c] *GRePS, Université Lyon 2, 5 avenue Pierre Mendès-France - 69676 BRON, France*
[d] *CATEL, PIBS-CP04-Bâtiment PIREN, Rue Henri Becquerel - 56038 Vannes Cedex*
[e] *IRSARP, Plaine Saint Gilles 72610 Saint Paterne, France*
[f] *TECHNOSENS, 15, rue d'Alsace, 69150 Décines, France*

Abstract

CIRDO project is intended to establish new healthcare systems to ensure the safety at home of seniors and people with decreasing independence. In particular, extending " e-lio footnote http://www.technosens.fr/ " devise, CIRDO aims to develop an audio/video based system which makes it possible for elderly to live with a sufficient degree of autonomy. To achieve this goal, generic purposes on video analysis and audio processing are discussed and implemented in the same process. Audio and video analysis algorithms are launched simultaneously and thinks to GPU implementation, the tasks are done in real-time. To comply with the requirements set out in the cirdo project namely: respect privacy and preserve personal data, the processing tasks is performed through background tasks, without any human intrusion. In order to have different types of fall for our experiences, multiple scenarios were designed and performed by several persons. Besides the technical aspect, the project also focuses on generic technology validation by potential users and their human environment (family, caregivers/care takers, ...). Psychological and ergonomic assessment on the use of services developed was conducted. It focused on the usefulness, usability and accessibility and acceptance of the tool. This evaluation was completed by critical investigation of knowledge acquired by professionals (geriatricians, caregivers school and associations).

Keywords: Ambient Assisted Living, Automatic recognition of aged voices, Scene analysis, Distress situation, User study

1. Introduction

Enabling seniors to stay in their own home as long as possible has become a major issue in developed countries for societal and economical motivations (improving well-being and autonomy, care and assistance). This concept, known as 'Ageing in place' has emerged due to the increase of life expectancy in all developed countries in the last decade. Indeed in the 2060s the number of people of 80 years old should be tripled [1]. In France, though the situation is less contrasted, 12 million people (out of 66 millions) received the Personalized Allocation of Autonomy (APA) [1], where 61% of them were at home and 39% were in specialized institutions [2]. When a senior is loosing autonomy, the assistance from a carer(s) becomes essential. This carer role is often played by relatives (partner, children...), who must ensure the care and often financial support. The challenge that our society must address is to enable seniors to age in place as autonomously and safely as possible while relieving some of the burden (stress, times, etc.) supported by the carers.

This is the aim of the Ambient Assisted Living (AAL) research programs which intend to develop services and innovative products that improve the quality of life of older people, maintain their independence and quality of life in a normal living environment whereas seniors are more affected by physical or cognitive diseases. This is in this context that our project, called CIRDO, is developed.

The aim of the project is to develop an ubiquitous system in the home, involving audio and video processing, making it possible to automatically detect falls and distress situations to call for help if necessary. The expected impact of such project is the improvement of the safety feeling of the person in loss of autonomy. By contrast with current worn system, the person does not need to wear sensors 24h/day. Figure 1 illustrates the installation and the use of the system inside the home of a person. Moreover, if the person falls in the camera range, the system can react even if the person losses consciousness or is dizzy. If the person is out of the camera range, the person still has the option to call for help thanks to the audio sensors. Thus, the proposed system overcomes some of the limitation of the currently available solutions.

[1]This allowance from the state is given to people over 60 years old in loss of autonomy

Several challenges must be addressed to achieve the project aims. Firstly, on a social aspect, though a large number of projects has been conducted in the domain, the fall process in the domestic setting is still difficult to model due to the variability of persons and home settings. Moreover, real acceptance of video/audio sensors is very variable according to the user's situation (does the intrusion worth the benefit ?). On the same line, not only the user's needs must be considered but also the feeling of their relatives and the professionals working in the home (house maiden, nurse, etc.). Secondly, on a computer science aspect, a number of issues must be addressed regarding video processing in varying light conditions, sound analysis with atypical users (aged voice) and noise, decision making with uncertain information (decision making from the inferences made by the video and audio processing), and Computer-Human interaction (does the system prompts the user and reacts properly?).

To tackle these issues, the project adopts a user-centred design [3] using a multi-disciplinary approach (partners includes computer scientists, social psychologists, field industrials, trainers of health professionals, members of a network of telehealth actors). The applied approach consisted in 1) modelling the fall process of aged people in their own home and to prospect the acceptability conditions of the CIRDO system by the users and their relatives 2) designing computational models on the basis of this study, 3) developing a complete monitoring system and 4) validating this system with the targeted users.

This paper, presents the CIRDO approach and the main results of the studies conducted so far which led to the development of a monitoring system not only compatible with the specific needs of the users but also transferable in their own psychosocial life environment. It is organised as follow, after a short state of the art presented in Section 2, Section 3 details the results of the field study which led the specification of the CIRDO system. The technical aspects of this system are described in Section 4. To build inference models, an experiment has been conducted in realistic settings, involving targeted users and users simulating an aged body. This experiment is described in Section 5. Finally, the paper ends with a discussion of the results obtained so far and outlines the future work.
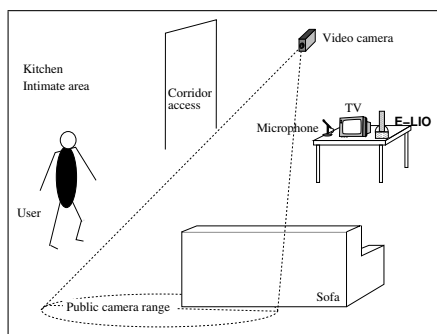

Figure 1: Global architecture of the CIRDO system.

2. State of the art

2.1. The effect of ageing

The ageing process is characterised by a decay of the main bio-physiological functions, affecting the social role and the integration of the ageing person in the society [4]. As a consequence, the person withdrawals from society to the home and her social and family role is weakened [5], but, above all, she losses goals and identity (disengagement theory [6]). However, a decrease of activities does not necessarily mean a disengagement in the remain ones [7]. Furthermore, through the ageing progression steps, these physiological and mental degradations are not irreversible. This progression can be normal, pathological, optimal or successful when the seniors succeed in adapting thereself to the changing situation. Thus, the emergence of disabilities is not originated only in the individual but also in the interaction with her environment. More precisely, it is the unfitness of the environment to the person entering into fragility that causes the incapacity situation [8]. In that case, two approaches exist in the literature [9]. An "integrative" approach which acts on the person (rehabilitation, devices...) and a "participatory" approach in which the environment is adapted to the person in loss of autonomy. The CIRDO project by detecting falls and other risky situations without constraining the person to wear a specific device, is clearly situated in the latter approach.

Apart from the social aspect, ageing also has effects on the voice and movements of the person. Indeed, aged voice is characterized by some specific features such as imprecise production of consonants, tremors and slower articulation [10]. From an anatomical point of view, some studies have shown age-related degeneration with atrophy of vocal cords, calcification of laryngeal cartilages, and changes in muscles of larynx [11, 12]. Thus, in the context of technology development aiming at helping aged people to live independently at home, the use of elderly voice can be an issue for performance of speech recognition systems. Some authors have reported that classical Automatic Speech Recognizers (ASR) systems exhibit poor performances with elderly voices because most acoustic models of ASR are acquired from non-aged voiced samples [13, 14]. Apart from the physical change of voice induced by ageing many other challenges are to be tackled in a home such as noise caused by devices, emotive voice. For more information about these issues the reader is refereed to [15].

2.2. The fall and its detection

In the literature, this domestic risky situation for an aged person is defined as: "inadvertently coming to rest on the ground, floor or other lower level, excluding intentional change in position to rest in furniture, wall or other objects" [16]. Falls are the cause of 80% of daily life injuries and 9,300 deads per year in the French population above 65 years old. About a quarter of the persons between 65 and 75 years old declared to have felt in the year [17]. It has been emphasized that the age, gender and health of the person can influence the gravity and the kind of injury [18].

Given the serious consequences that can follow a fall, a number of projects have developed solutions to detect it as early as

possible. In most cases, the authors propose the use of a worn sensor [19]. A few projects have considered microphones to detect fall. It is often fused with other modalities like motion thanks to the use of an accelerometer [20, 21]. Doukas et al. [21] proposed a patient awareness system to detect body fall and vocal stress in speech expression through analysis of motion data and acoustic, but in this case the person is constrained to wear sensors continuously. Another approach proposed a dialogue system in order to replace traditional emergency systems [22]. However, the vocabulary of the prototype was limited (yes or no) and the system was not tested with aged users. Another popular approach is the use of companion robots like in the CompanionAble project [23]. The companion robot is able to move nearby the person so that the user is close to the sensor. The robot is able to recognize commands uttered in Dutch, no specific application to distress sentence recognition were reported by the authors. This system is limited by the capacity of the robot to move close to the person and we are not aware of studies reporting the transferability of such devices in the own home of the person. Recent studies have shown that audio sensor, though being intrusive, are accepted by the elderly population [24][25].

The originality of the Cirdo project lies in its combination of audio and video analyses together with a deep field study. Video processing is highly accurate to detect movements of the person but its action is limited to the camera range and the light condition while microphones can sense sounds outside the camera range and permits voice recognition but cannot detect person movement and is challenged by noise.

3. Field studies

As previously said, the fall process in the ageing people is often the conjunction of several "risky" factors of intrinsic type (related to fragility, people own apprehension...) and extrinsic (environment) [26] as well as sociocultural effects (personal or shared experience of the fall, life history...) [27].

The objective of the field study is to analyse in depth the different fall cases in order to identify the key poses (behaviour, gestures and poses during the fall) and the key phrases (speech, sounds) characterising the risky situations that can be used to construct computational models. To do this, a number of questions guided the analysis: what are the daily living activities (ADL) performed at home by the seniors? What are the characteristics of the falls and the causing factors? On which descriptors the Cirdo system can be based to detect automatically these falls? What are the main priorities for the inclusion of the system both from a sensor point of view and a dissemination point of view (respect for private life vs. risky areas in the house)?

To do this, various methods have been deployed: semi-structured interviews with 65 aged people, analysis of the activities of daily life conducted at seniors homes, analysis and script falls methods (recorded simulation and observations of falls and Personas method). From these, 12 "scripts" of typical falls have been constructed by describing on the one hand the fall context (person, activity been performed, location, moment, causes...) and on the other hand the members of the body and

their movements during the fall as well as the person's reaction once on the floor (try to get back on her feet, crawling...) and the duration of the immobilisation. Figure 2 shows a script of a typical fall. Phrases uttered during the fall have also been identified "Aahhh, zuuut, qu'est-ce qui m'arrive? Oh merde, merde!".
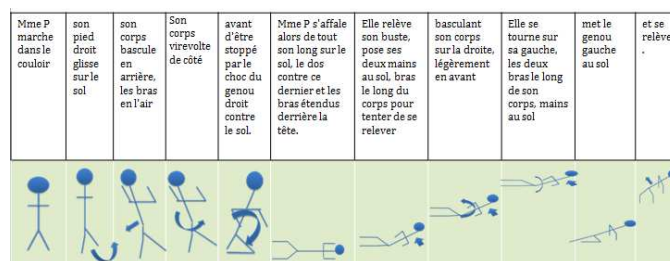

Figure 2: Different phases of a fall

Using the Personas method, we defined scenarios which describe the context and the process of the falls to:

1. obtain a clear representation of the 28 identified falls so that a categorisation in 3 disjoint classes: falling, slipping and stumbling, was possible;
2. ask 20 participants to play a fall respecting these scenarios in order to record them in the controlled conditions of a living lab.

Figure 3 describes an example of fall scenario obtained thanks to this method.

```
"She gets up in the middle of the night to drink
a glass of water in the kitchen. She is walking
in the dark when suddenly her foot slips on the
floor. She loses her balance and her whole body
swings backward. She exclaimed 'ouaahhhhhhh!!!'
Her body turns back on the right side. Her right
knee hits first the ground violently, then her
whole body. She finds herself on the back, arms
extended behind head."
```

Figure 3: Fall scenario using the Personas method

4. The CIRDO system

4.1. Global Architecture

The different components of the Cirdo system and their connection are described Figure 4. The video camera stream is processed by the Video module which extracts continuously the main positions and situations of the person as well as detect falls. The audio stream is processed by the Audio module which extracts continuously the speech and sound utterances in the home. These events are sent to the Fusion/Decision module which analyses the incoming events, detects risky situations and decides whether an emergency call must be emitted through the *e-lio* system (see http://www.technosens.fr/).
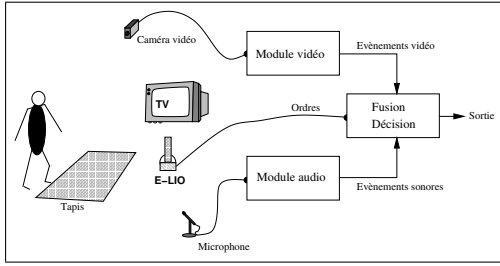
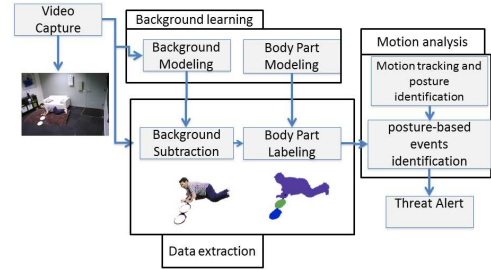Figure 4: Global architecture of the CIRDO system.

Figure 5: Video analysis : the overall system

4.2. Video analysis

Video capture in a private indoor space can benefit from some favorable operating conditions: not subject to vibration nor direct weather conditions. However, the daily living environments face specific difficulties which make the video analysis and the scene interpretation a hard task to accomplish. Among the most significant are: (i) changes in light levels between bright and dark areas variations caused by artificial light sources, (ii) sudden changes of lighting (turning on or off the indoor lights, a curtain that moves), (ii) foreground colors and patterns are close to background one (the problem of camouflage) as well as the lack of contrast between home furnishings, (iii) the variety of light sources, often artificial, which can generate shadows connected to objects and affect detection and identification of the subject of interest, (iv) the plethora of objects (furniture, ornaments, carpets, etc.) in the environment (living room, kitchen, bedroom, etc.) and some of which are likely to be moved [28].

CIRDO is built in order to accomplish several tasks for monitoring the safety of elderly in home environments. The vision modules is responsible for both locating individuals present in the scene, tracking their movement and identifying some unusual events like falling, or a long period of inactivity. Given the application context, the cameras are static. The camera was then calibrated to determine its intrinsic parameters and extrinsic parameters with respect to the plane of the ground. OpenCV vision library provides efficient system calibration algorithms. Area inside the camera's field of view is as well estimated.

In order to comply with a strict code of ethics that fully protects user privacy and ensure an efficient detection, the video stream is acquired continuously and the process runs in background. The process chain is performed in automatic way. This means that the process runs without being connected to any server and doesn't require any operator intervention.

In this context we are interested in the detection of risky situations; with focus on detecting several types of fall. To this aim, as mentioned above we built scenarios simulating real life situations. Our video processing stage is structured as Figure 5. It based on the extraction of the silhouette vs. the separation between foreground and background. The extracted silhouette is used to obtain useful characteristics for detection of movements events "at risk". Tracking the movement of this figure allows us to identify a fall. The treatment consists of four steps: (1) learning the background, (2) extraction of the silhouette, (3) analysis of posture and motion tracking, (4) distress posture/fall

recognition and generating alert tag sent to the audio-video fusion process. Moreover, as with the audio processing, to respect the privacy of the person, the scenes that are not related to distress are removed. The first step performs background learning and subtraction for extracting the silhouette, which should be processed for identification or tracking. The difficult part of background subtraction is not the differencing itself, but the maintenance of a background model [29] especially in case of moved objects, sudden or progressive changes of illumination or any other alterations of the appearance of the background. The model used is a variant of Stauffer and Grimson [30] method called Mixture Of Gaussian or MOG adapted for multi-level representation: pixel, region and image. On the first level, we use an enhanced Mixture of Gaussians, built on both chrominance-luminance and chrominance-only spaces, which handles global illumination changes [31].

On the second level, we improve segmentation results, in interesting areas, by using statistical foreground models updated by a high-level tracking of body parts. Each body part is represented with a set of template characterized by a feature vector built in an initialization phase. To improve tracking and identification of the posture, each body part is represented by a blob model. The approach starts with three body parts: head, upper part and lower part. Then analysis of upper and lower part adds new body parts when it is possible: arms and legs. Finally, high level tracking is done by finding blob-template correspondences via distance minimization in feature space. Correspondences are then used to update foreground models, and a graph cut algorithm [32], which minimizes a Markov random field energy function containing these models, is used to refine segmentation and recognition [33]. This step is reinforced by detecting and tracking the position of the user [34]. To cope with the problem of real-time, the video processing module is improved with NVIDIA GPU cards with CUDA technology.

4.3. Audio module: automatic aged speech recognition

The audio module of the CIRDO project, CirdoX, is in charge of 1) detecting sound events; 2) discriminating between noise and speech and 3) performing aged speech recognition.

The module has been designed to operate as a background task with complete autonomy. It doesn't require any action from the elderly for initialization, settings and configuration. The system is continuously listening to the sound environment to recognize and detects any occurrence of speech that can be emitted during a distress situation. Anything that is not distress
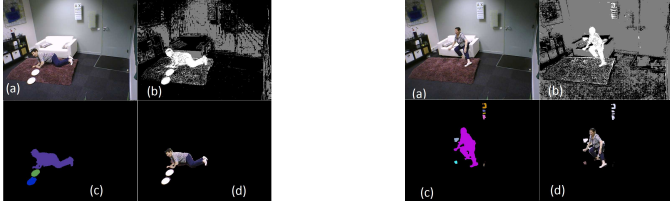
Figure 6: distress situation when the person can not sit up (left sequence), Video analysis result: fall detection (left sequence) (a) initial record, (b) discrimination between the background and foreground, (c) extraction of silhouette (d) situation identification

speech is rejected in order to preserve the privacy of the elderly. The system is operating in pseudo real time to trigger an alarm immediately after the recognition of a distress call. This service is intended for the elderly, it should therefore be adapted to the voice particularities of this part of the population. The system is designed for a user generic older voice model, without the need for an adaptation to speaker individually. Also, the stakes are that it is robust to the emotions when distress situations are occurring.

For enabling adaptation to aged speakers a large amount of aged speech is needed. However, there is no such corpus available in French. We thus performed our own data collection of French aged voice. However, this collection was insufficient to learn acoustic models from scratch. That is why we used acoustic models learned on typical speaker that we adapted using the corpus of aged voices. With this technique, the average Word Error Rate (WER) decreased to 14.52% compared with the 43.47% WER without adaptation. From an applicative point of view, this test shows that even though the size of the targeted user corpus is small, we can reach sufficiently good performance to use the audio module in Cirdo. For more details about the study, the reader is refered to [35].

After a bibliographical study in collaboration with the GRePS laboratory and in the prolongation of previous studies [36], we defined a list of home automation orders (intentional calls to the system) and of distress calls that the person could utter during a distress situation to request for assistance. This list was completed with casual sentences. At the beginning of the Cirdo project, a survey was conducted by the GRePS laboratory [37] in order to determine what sort of sentences were truly uttered by aged people in distress case and during a fall and the collected sentences were added to the list. Ten samples of each kind (distress, home automation order, casual) are given in Table 1. To enable comparison of system performance between elderly and non-elderly, the corpus contains recordings made by these two groups of people. This corpus is called the AD80 corpus (Anodin-Détresse: *anodin* means casual and *détresse* means distress). The non-aged part of the corpus was previously recorded in our laboratory in 2004 and was complemented in 2013 with sentences based on [37]. The collection of the aged part of the AD80 corpus was performed sporadically from 2009 to 2012 in collaboration with a rehabilitation centre, volunteers and a nursing home. Targeted speakers were persons aged of more than 60 years old. The recording was done with a single microphone positioned about 30 cm from the speaker's

mouth. The recording was done using a computer and a home made software to prompt sentences to be read by the speaker and to record the utterances using voice activity detection.

The Automatic Speech Recognition (ASR) toolkit implemented in the module was Sphinx3 [38]. This decoder used a context-dependent acoustic model with 3-state left-to-right HMM. The acoustic vectors are composed of 13 MFCC coefficients, the delta and the delta delta of each coefficient.

A general language model (LM) was estimated from French newswire collected in the *Gigaword* corpus. It was 1-gram with 11,018 words. Moreover, to reduce the linguistic variability, a 3-gram domain language model was learnt from the sentences used during the corpus collection, with 88 1-gram, 193 2-gram and 223 3-gram models. Finally, the language model was a 3-gram-type which results from the combination of the general language model (with 10% weight) and the domain one (with 90% weight). This combination has been shown as leading to the best WER for domain specific application [39]. The interest of such combination is to bias the recognition towards the domain LM but when the speaker deviates from the domain, the general LM makes it possible to correctly recognise the utterances.

To discriminate distress sentences from other sentences (e.g., personal communications) each ASR hypotheses are compared with the list of distress sentences at a phonetical level (see Table 1) using the Levenshtein distance. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other.

In order to assess the filtering, we realized a decoding with the 43 elderly speakers from the AD80 corpus, including 2796 distress/home automation sentences and 3006 casual sentences, for a total duration of 2 hours 12 minutes. The casual utterances were used as disrupter, with some sentences far from the distress ones: for instance *Les patates sont cuites* (Potatoes are cooked), or closer, for instance *Le médecin a appelé* (The doctor called). The average WER with the adapted acoustic model was 14.5% for the distress sentences, and was much higher for the casual sentences, 87.5%, due to the adapted-to-distress language model. At the point of equal error, we obtained the positive and negative test showed in table 2, and recall, precision and F-measure equal to 88.4, 86.9 and 87.2% .

Table 2: *F-measure of the distress sentence detection task.*

Threshold = 0.7	Distress	Casual
Distress	TP = 2472	FP = 374
Casual	FN = 324	TN = 2632

4.4. Data fusion/Decision module

The data fusion module is in charge of sending an alarm when the video module detects a fall or when the last stage of the audio module identifies a distress call or an home automation order. This module acquires continuously the events sent by the video and audio modules and maintains the knowledge

Sample	Distress Sentence	Home Automation Order	Casual Sentence
1	Qu'est-ce qui m'arrive !	Appelle quelqu'un e-lio !	Bonjour madame !
2	Oh là !	e-lio, appelle quelqu'un !	Ça va très bien.
3	Oh là ! Je saigne ! Je me suis blessé !	e-lio appelle les secours !	Ce livre est intéressant.
4	Aïe ! J'ai mal !	e-lio appelle ma fille !	Il fait soleil.
5	Je peux pas me relever !	e-lio, appelle du secours !	J'ai ouvert la porte.
6	Aidez-moi !	e-lio appelle le SAMU !	Je dois prendre mon médicament !
7	Au secours !	e-lio appelle les pompiers !	J'allume la lumière !
8	Je me sens mal !	e-lio appelle mon fils !	Je me suis endormi tout de suite !
9	Je suis tombé !	e-lio tu peux téléphoner au SAMU ?	Le café est brûlant !
10	Du secours s'il vous plaît !	e-lio tu peux appeler une ambulance ?	Où sont mes lunettes ?

Table 1: Examples of sentences of the AD80 corpus

of the situation to provide context for the decision. Variables such as the number of persons in the camera range, the luminosity of the room and the audio energy in the room can enable the characterisation of the situation. For instance, when the luminosity in the room is low, the information coming from the video module will be less reliable so that decision would mostly be based on audio processing. Once a distress event is identified, a message is sent to the communication module that will handle the emergency call.

4.5. Communication module

The communication module is actually a commercial product named *e-lio* made by the Technosens company. In the house of the elderly person, this equipment, composed of video camera, a central unit and a 3 buttons remote control which is a phone as well, is connected to the user's televisions and gives her access to a lot of connected TV services such as: television, phone calls, video calls, radio, pictures sent by family, games. The system has also an emergency mode that handles the connection with a relative till it reaches a carer (in case the first contact did not answer the call).

5. Experiments for Corpus Collection

To develop the algorithms needed for the audio and video automatic detection of falls in the home, experiments were conducted in a smart room configured to look like a standard room (chairs, carpet, coffee table, TV. . .). These experiments aimed at collecting significant number of samples of data to learn and test the algorithms by asking targeted users to replay the scenarios that were defined in the field study (cf. Section 3).

5.1. Experimental protocol

The protocol was as follows. Each participant was introduced to the context of the research and was invited to sign a consent form. The participants played five types of fall chosen from the 28 risky situations identified (1 slip, stumble 1, 2 falls in a stationary position and a position of hip blocked on the sofa). Figure 7a shows a excerpt of a participant simulating a fall. These situations were selected because they were representative falls in the home and because they could safely be
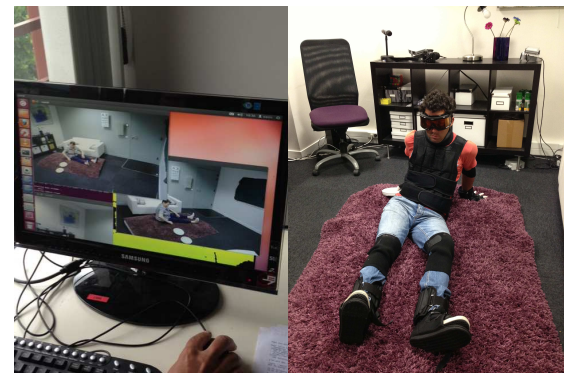
(a) Video capture   (b) Simulator

Figure 7: Snapshots of the experiment.

played by the participants. Two other scenarios, called "true-false", were added to challenge the automatic detection of falls. The first "true-false" consisted in rapidly picking up magazines on the floor (close to a situation of fall) while the second was to try to catch a remote control on a coffee table when the person is sitting on the sofa (close to the situation in which the person has a blocked hip).

Before each scene to play, the experimenter explained the scene to the participant specifying the context (location, moment of the day. . .), the activity (what the person was being to do and what she wanted to do when the incident occurred) and the gestures and words produced before, during and after the fall. Then, the person was asked to repeat several times the scene before starting the camera and microphone recording. In average, the experiment lasted 2h30 per person.

5.2. Materials

The experimental room and the equipment used are depicted Figure 8.

Regarding audio analysis, we conducted a multi-source capture with two wireless Sennheiser microphones. A eW-300-G2 type ME2 microphone was placed in the suspended ceiling of the experiment room, and a SKM-300-G2 microphone was placed on a furniture close to the subject. The recording was
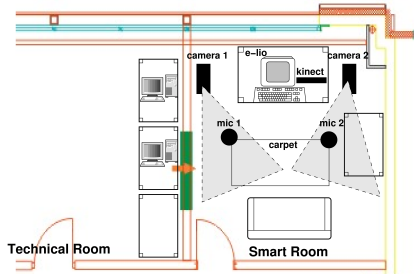
Figure 8: Experimental room and equipment.

No	Age/Sexe	Occurence number of each scenario							Time
		F1	F2	F3	F4	B	TN1	TN2	
Y01	30/M	1	2	1	1	1	1	1	8mn 40s
Y02	24/F	1	1	1	1	1	1	1	4mn 35s
Y03	29/M	1	1	1	3	1	1	2	6mn 30s
Y04	44/M	1	1	1	2	1	1	1	5mn 54s
Y05	16/M	1	1	1	3	2	1	2	8mn 50s
Y06	16/M	1	1	1	1	1	1	1	5mn 07s
Y07	52/M	1	1	1	1	1	1	2	5mn 17s
Y08	28/M	1	1	2	1	1	1	2	7mn 04s
Y09	52/F	3	1	1	1	1	1	2	6mn 48s
Y10	23/M	1	1	2	3	1	1	1	5mn 50s
Y11	40/F	2	1	2	3	1	1	1	7mn 31s
Y12	40/F	1	1	1	2	2	1	2	8mn 01s
Y13	25/F	2	2	1	1	1	1	1	5mn 54s
A01	83/F	1	1	3	2	3	2	2	9mn 07s
A02	64/F	1	2	1	2	1	1	2	6mn 31s
A03	61/M	1	1	1	3	1	1	1	6mn 00s
A04	66/M	1	1	2	1	2	2	2	7mn 16s
ALL		21	19	23	31	22	19	26	1h 55mn

Table 3: Audio/Video Corpus for the two group of participants

performed in the technical room by a computer connected to the microphones high-frequency receivers thanks to a National Instruments PCI-6220 8-channel card. A loudspeaker in the control room enabled experimenters to hear the progression of the scene in the smart room. The National Instruments PCI-6220 card was controlled by the StreamHIS software. This software enables data acquisition and was developed by the GETALP team.

Regarding video analysis, we conducted the video capture with two cameras (webcam Sony PSEye: 640x480 60Hz connected by a USB USB 2.0) fixed to the wall of the smart room in front of the sofa and pointing towards it, and a depth camera (Microsoft Kinect). Synchronization and recording of video streams were performed in the technical room, by a computer Intel i3 3.2GHz, 4GB RAM, NVidia GeForce GTS 450 512MB. The tools enabling video acquisition was developed within the team SAARA (based the OpenCV library). Video processing (background learning and subtraction, body parts segmentation) is performed on a PC cluster, taking advantage of algorithms parallelization on CPU and GPU. Motion tracking and identification of hazardous situations is done on a PC (Intel i7 3.5GHz, 4GB RAM, NVIDIA GTX 660 4GB GeFore).

5.3. Corpus

The targeted participants were elderly person that were still able to play the scenario defined in Section 3. However, the recruitment of such population is a long task. So, given that the aim was mostly to collect data that were realistic but not necessary real, people under 60 were recruited invited to wear an equipment (see Figure 7b) which hampered their mobility and reduced their vision and hearing to simulate aged physical conditions.

Overall, 17 participants were recruited (9 men and 8 women) with mean age of 40 years old (SD 19.5). Among them 13 people were under 60 and worn the simulator. The persons of the aged group were between 61 and 83 years old (mean 68.5). The details relative to the corpus composition are given Table 3. Scenarios are relative to fall (F1 to F4), blocked hip (B) and True Negative (TN). The necessary time for achieving one scenario was twice for the aged group.

In order to make the corpus exploitable, the team is currently annotating the distress situation in the data. This is made using the Advene software [2] developed at the LIRIS laboratory. Ad-

vene allows to organize annotations elements — such as type of annotations, annotations, and relationships — under schemes defined by the user. In this work, a unique schema is used which contains the following annotation types: fall period and distress situation, and now we plan to extend it to include information about video and speech features.

6. Conclusion

This paper summarizes the research results achieved within CIRDO project which is focused on an innovative tools to facilitate safety of elderly living in domestic environments and caregivers. The project brought together several partners which are actually complementary, involving multidisciplinary studies, including sensors acquisition, audio processing, video analysis and humanities. In particular, the proposed solution have tackled both the algorithmic and architectural technologic solution and taking into account the human factors (acceptability, ethic). The aim of CIRDO project is to simplify access to equipment daily taking into account the needs and uses of elderly people. Studies conducted by the partners from humanities allow CIRDO to establish a link between the needs, people and technology, and proposes solutions by gathering knowledge from different domain.

The CIRDO project's key success is combining engineering sciences and the humanities to propose novel solution to help elderly people and persons with impaired. On the technical side, CIRDO proposes an original solution which relies on a fusion of audio and video data. In order to respect privacy and preserve personal data the processing is achieved without any external human intervention. The developed system will be integrated into the *e-lio* service platform.

To deal with real time constraint the video processing is done on GPU, allowing quick reaction to distress situation and in particular the fall cases. Regarding audio analysis, we focused initially on the recognition of the older voices, but affects also

---

[2] http://liris.cnrs.fr/advene/

have an influence on the performance of the ASR [40]. Further studies are still ongoing and we are currently recording a corpus adapted to this situation.

additional information

- Conflict of interest: none

References

[1] Eurostat, Projections de population 2008-2060, Communiqués de presse Eurostat (26 août 2008).

[2] V. Bellamy, C. Beaumel, Bilan démographique 2012. la population croît, mais plus modérément, INSEE Première 1429 (2013) 1–4.

[3] D. A. Norman, S. W. Draper, User Centered System Design; New Perspectives on Human-Computer Interaction, L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1986.

[4] J. Ogg, C. Bonvalet, L'état des enquêtes sur l'entraide en europe, Rapport final de recherche CNAF, Mire, Collections de l'INED (2006).

[5] I. Rosow, Socialization to old age, University of California Press, Berkeley and Los Angeles, California, 1974.

[6] E. Cumming, W. E. Henry, Growing Old: The Process of Disengagement, Basic Books, New York, 1961.

[7] S. Clément, M. Membrado, Penser les vieillesses. Regards anthropologiques et sociologiques sur l'avancée de l'âge, Paris : Seli Arslan, 2010, Ch. Expériences du vieillir : généalogie de la notion de déprise, pp. 109–128.

[8] A. Newel, P. Gregor, User Sensitive inclusive Design, in: Actes du Colloque Interaction Homme Machine et Assistance, Metz, France, 2001, pp. 18–20.

[9] E. Ebersold, Le champ du handicap, ses enjeux et ses mutations : du désavantage à la participation sociale, Handicap (2002) 149–164.

[10] W. Ryan, K. Burk, Perceptual and acoustic correlates in the speech of males, Journal of Communication Disorders 7 (1974) 181–192.

[11] N. Takeda, G. Thomas, C. Ludlow, Aging effects on motor units in the human thyroarytenoid muscle, Laryngoscope 110 (2000) 1018–1025.

[12] P. Mueller, R. Sweeney, L. Baribeau, Acoustic and morphologic study of the senescent voice, Ear, Nose, and Throat Journal 63 (1984) 71–75.

[13] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, K. Shikano, Acoustic models of the elderly for large-vocabulary continuous speech recognition, Electronics and Communications in Japan, Part 2 87 (2004) 49–57.

[14] R. C. Vipperla, S. Renals, J. Frankel, Ageing voices: The effect of changes in voice parameters on ASR performance, EURASIP Journal on Audio, Speech, and Music Processing (2010) 1–10.

[15] M. Vacher, F. Portet, A. Fleury, N. Noury, Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges, International Journal of E-Health and Medical Communications 2 (1) (2011) 35–54.

[16] WHO global report on falls prevention in older age, Tech. rep., World Health Organization (2007).

[17] INPES, Mieux prévenir les chutes chez les personnes âgées, La santé de l'homme, Institut National de Prévention et d'Education pour la Santé, 381 (2006).

[18] C. J. Todd, C. Ballinger, S. Whitehead, A global report on falls prevention: reviews of socio-demographic factors related to falls and environmental interventions to prevent falls amongst older people living in the community, Tech. rep., Geneva: World Health Organization (2007).

[19] F. Bloch, V. Gautier, N. Noury, J. Lundy, J. Poujaud, Y. Claessens, A. Rigaud, Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects 54 (2011) 391–398.

[20] M. Popescu, Y. Li, M. Skubic, M. Rantz, An acoustic fall detector system that uses sound height information to reduce the false alarm rate, in: Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008, 2008, pp. 4628–4631.

[21] D. Charalampos, I. Maglogiannis, Enabling human status awareness in assistive environments based on advanced sound and motion data classification, in: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments, 2008, pp. 1:1–1:8.

[22] M. Hamill, V. Young, J. Boger, A. Mihailidis, Development of an automated speech recognition interface for personal emergency response systems, Journal of NeuroEngineering and Rehabilitation 6 (1) (2009) 26.

[23] P. Milhorat, D.Istrate, J. Boudy, G. Chollet, Companionable Hands-free speech-sound, in: EUSIPCO (European Signal Processing Conference), The European Association for Signal Processing (EURASIP), Bucarest, Romania, 2012, pp. 1678–1682.

[24] F. Portet, M. Vacher, C. Golanski, C. Roux, B. Meillon, Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects, Personal and Ubiquitous Computing 17 (2013) 127–144.

[25] R. Vipperla, M. Wolters, S. Renals, Spoken dialogue interfaces for older people, in: K. J. Turner (Ed.), Advances in Home Care Technologies, Vol. 1, IOS Press, 2012, pp. 118–137.

[26] C. Ballinger, S. Payne, The construction of the risk of falling, among and by older people, Ageing and Society 22 (3) (2002) 305–321.

[27] M.-E. Bobillier-Chaumon, C. Michel, F. Tarpin-Bernard, B. Croisile, Can ict improve the quality of life of elderly adults living in residential home care units? from actual impacts to hidden artefacts, Behaviour & Information Technology ((to appear)) 1–17.

[28] R. Cucchiara, C. Grana, A. Prati, R. Vezzani, Computer vision system for in-house video surveillance, Vision, Image and Signal Processing, IEE Proceedings - 152 (2) (2005) 242–249.

[29] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: principles and practice of background maintenance, in: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, Vol. 1, 1999, pp. 255–261. doi:10.1109/ICCV.1999.791228.

[30] C. Stauffer, W. E. L. Grimson, Adaptive background mixture models for real-time tracking, in: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., Vol. 2, 1999.

[31] R. Deeb, E. Desserée, S. Bouakaz, Real-time two-level foreground detection and person-silhouette extraction enhanced by body-parts tracking, in: Proc. SPIE, Vol. 83010R, 2012, pp. 1–8. doi:10.1117/12.908435.

[32] C. Rother, V. Kolmogorov, A. Blake, "grabcut": Interactive foreground extraction using iterated graph cuts, in: ACM SIGGRAPH 2004 Papers, SIGGRAPH '04, ACM, 2004, pp. 309–314.

[33] R. Deeb, D. Elodie, S. Bouakaz, Extraction de silhouette et suivi des parties du corps dans un environnement non contrôlé, in: CORESA 2012, 2012.

[34] R. Deeb, F. Lédée, E. Desserée, S. Bouakaz, Méthode robuste pour la détection de chute dans un environnement non contrôlé, in: RFIA 2012 - RFIA 2012 - Atelier VISAGES oai:hal.archives-ouvertes.fr/, 2012.

[35] F. Aman, M. Vacher, S. Rossato, F. Portet, Analysing the Performance of Automatic Speech Recognition for Ageing Voice: Does it Correlate with Dependency Level?, in: 4th Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France, 2013, pp. 9–15.

[36] M. Vacher, P. Chahuara, B. Lecouteux, D. Istrate, F. Portet, T. Joubert, M. Sehili, B. Meillon, N. Bonnefond, S. Fabre, C. Roux, S. Caffiau, The SWEET-HOME Project: Audio Technology in Smart Homes to improve Well-being and Reliance, in: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13), Osaka, Japan, 2013, pp. 7298–7301.

[37] M.-E. Bobillier-Chaumon, B. Cuvillier, S. Bouakaz, M. Vacher, Démarche de développement de technologies ambiantes pour le maintien à domicile des personnes dépendantes : vers une triangulation des méthodes et des approches, in: Actes du 1er Congrès Européen de Stimulation Cognitive, Dijon, France, 2012, pp. 121–122.

[38] K. Seymore, C. Stanley, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, E. Thayer, The 1997 CMU Sphinx-3 English broadcast news transcription system, DARPA Broadcast News Transcription and Understanding Workshop.

[39] B. Lecouteux, M. Vacher, F. Portet, Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions, in: Interspeech 2011, Florence, Italy, 2011, pp. 1–4.

[40] F. Aman, V. Auberge, M. Vacher, How affects can perturbe the automatic speech recognition of domotic interactions, in: Workshop on Affective Social Speech Signals, Grenoble, France, 2013, pp. 1–5.