The CIRDO Corpus: Comprehensive Audio/Video Database of Domestic Falls of Elderly People

Michel Vacher, Saida Bouakaz, Marc-Eric Bobillier-Chaumon, F Aman, Rizwan Ahmed Khan, S Bekkadja, François Portet, Erwan Guillou, S Rossato, Benjamin Lecouteux

The CIRDO Corpus: Comprehensive Audio/Video Database of Domestic Falls of Elderly People

M. Vacher, S. Bouakaz, M.-E. Bobillier Chaumon, F. Aman, R. A. Khan, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, B. Lecouteux

CNRS, LIG, F-38000 Grenoble, France
Univ. Grenoble Alpes, LIG, F-38000, Grenoble, France
LIRIS, UMR 5205 CNRS/Université Claude Bernard Lyon 1, F-69622 Villeurbanne, France
GRePS, Université Lyon 2, F-69676 Bron, France
Michel.Vacher@imag.fr, saida.bouakaz@univ-lyon1.fr, marc-eric.bobillier-chaumon@univ-lyon2.fr

Abstract

Ambient Assisted Living aims at enhancing the quality of life of older and disabled people at home thanks to Smart Homes. In particular, regarding elderly living alone at home, the detection of distress situation after a fall is very important to reassure this kind of population. However, many studies do not include tests in real settings, because data collection in this domain is very expensive and challenging and because of the few available data sets. The CIRDOcorpus is a dataset recorded in realistic conditions in DOMUS, a fully equipped Smart Home with microphones and home automation sensors, in which participants performed scenarios including real falls on a carpet and calls for help. These scenarios were elaborated thanks to a field study involving elderly persons. Experiments related in a first part to distress detection in real-time using audio and speech analysis and in a second part to fall detection using video analysis are presented. Results show the difficulty of the task. The database can be used as standardized database by researchers to evaluate and compare their systems for elderly person's assistance.

Keywords: audio and video data set, multimodal corpus, natural language and multimodal interaction, Ambient Assisted Living (AAL), distress situation.

1. Introduction

Twenty first century is witnessing a rapid growth in population over 65 years old worldwide (Who, 2013). This change in demography put pressure on governments which are incurring unprecedented expenditure to support ageing population since ageing is correlated with an increase in health and daily living support. For instance, in France, 12 million out of 66 million people are receiving the personalized allocation of autonomy (PAA)[1], a government financial aid for elderly person to support them in their daily life. This demographic change is very challenging to the society, governments and to technologists to come up with reliable and sustainable solutions to help ageing population to live as independently and safely as possible while relieving their family's emotional and financial burden. However this trend also provides the opportunity to come up with technologies that can help ageing population to carry on with their lives without compromising their privacy and at the same time provide assistance to caregivers i.e. nurse, partner, children etc.

One of the main sources of stress in the ageing population for which technology can be of help is the fall incident. Indeed fall is a major source of injury for elderly people, between 20 and 40 percent of adults over 65 who live at home fall (Saim, 2014). Fall could lead elderly persons to faint or become unconscious. Another emergency situation is the inability to call for a help in distress situation of some aged persons due to their reduced mobility. In the absences of another person in the home (such as a carer), these situations could have disastrous effect.

Several studies had worked on an ICT solutions to bring autonomy and security back to the user (Popescu et al., 2008; Hamill et al., 2009; Young and Mihailidis, 2010; Bloch et al., 2011). These studies have followed the paradigm shift in Human Computer Interaction design, from computer-centred designs to human-centred designs (Pantic et al., 2006). Thus reliable and robust analysis of ageing population's gesture and voice has become inevitable in many application areas, such as social robots, emergency situation detection, behaviour monitoring, communication support, etc. Since many of these studies are based on statistical analysis (either for modelling or for machine learning), there is a dire need of databases of audio and video recordings to build and evaluate performance of systems that can analyse ageing population's gesture and voice. Indeed evaluation and benchmarking of systems / algorithms must be performed on standardized and accessible databases for meaningful comparison. In the absence of comparative tests on such standardized databases it is difficult to exhibits the relative strengths and weaknesses of these systems.

To the best of our knowledge there exists no such database that has video with audio recordings of aged persons during fall and distress situations while it is recognised as one of the major problem in the evaluation of fall detectors (Igual et al., 2013). Young and Mihailidis (Young and Mihailidis, 2013) have collected a corpus of spontaneous speech, read sentences, and emergency scenarios from adult actors aged 23–91 years but it is in English and do not contain videos recording.

This could be due to the fact that it is very difficult to setup a camera and microphone to record such events. Secondly, there is an ethical issue associated with recording aged person in distress situation and later making it available to public. Very few studies were able to record real falls of elderly persons (Kangas et al., 2012; Klenk et al., 2011) but those were only concerned with accelerometer sensors. To overcome this bottleneck in the research and development

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

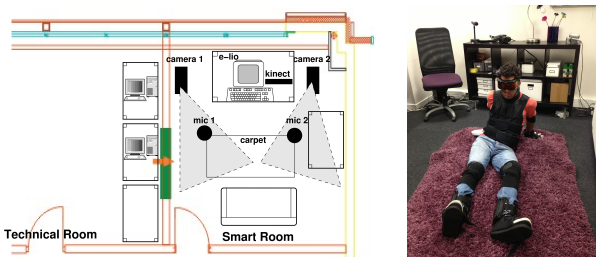

Figure 1: (a) Experimental room of the LIG with its equipment and (b) screen shot of a young participant wearing a simulator during the training phase.

Reference	Age	Sex	Reference	Age	Sex
Y01	30	M	Y07	52	M
Y02	24	F	Y08	28	M
S01	83	F	S04	66	M
Y03	29	M	Y09	52	F
S02	64	F	Y10	23	M
S03	61	M	Y11	40	F
Y04	44	M	Y12	40	F
Y05	16	M	Y13	25	F
Y06	16	M			

Table 1: Characteristics of the participants of the records (Sxx: senior people - Yxx: younger people)

of emergency situation detectors, we introduce the CIRDO database which is part of the CIRDO project (Bouakaz et al., 2014) whose aim was to develop an audio/video emergency detection system for elderly persons.

The CIRDO database contains video with audio (French language) recordings of aged persons during falls and distress situations. CIRDO database can be used as standardized database by researchers to evaluate and compare their systems for elderly person's assistance. In this article, we provide all the details related to various steps that have been carried out to record data in such a challenging situation.

2. Data Acquisition

Recording sessions were based on written scenarios. These scenarios were well elaborated thanks to a field study involving 15 elderly persons (Bobillier-Chaumon et al., 2012). Four scenarios were related to fall (F1 to F4), one to blocked hip (B) and two were true negative (TN1 and TN2). Recording sessions were conducted in the DOMUS smart room of the LIG laboratory configured to look like a standard room (chairs, carpet, coffee table, TV...).

2.1. Participants

Ideally scenarios should be played by elderly people. However, it is cumbersome to find elder person who is capable and willing to play such scenarios (refer Section 2.2.). To record realistic data (but not necessarily played by elder person), people under 60 were recruited. Then, these volunteers were instructed to wear an equipment i.e. *old age simulator* (see Figure 1). Old age simulator hampered mobility, reduced vision and hearing. Overall 17 participants were recruited (9 men and 8 women) with mean age of 40 years (SD 19.5). Among them 13 people were under 60 and worn the simulator.

Figure 2: Excerpt of 2 video records showing each the fall of an elderly participant.

2.2. Experimental Protocol

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 28 risky situations identified as: slip, stumble, falls in a stationary position and a position of hip blocked on the sofa. Figure 2 shows two elderly participants when they fall on the carpet. These situations were selected because they were representative falls in domestic environment and could be played safely. Two other scenarios, called "true-false", were added for the evaluation and verification of automatic fall detection. The first "true-false" situation consists of a rapid action of picking up of magazines from the floor (close to a situation of fall), while the second was to try to hold a remote control from the coffee table when the person is sitting on sofa (close to a situation in which the person has a blocked hip).

Scenarios included call for help sentences. These sentences were chosen based on sentences identified during the field study and on sentences extracted from the AD80 corpus (Aman et al., 2013) which was built with the aim of elderly voice study. Table 2 gives some examples.

Distress Sentence	AD80 Sentence
Aïe aïe aïe	Aidez-moi
Oh là	Au secours
Merde	e-lio, appelle du secours
Je suis tombé	e-lio appelle ma fille
Je peux pas me relever	Appelle quelqu'un e-lio
Qu'est-ce qu'il m'arrive	e-lio, appelle quelqu'un
Aïe ! J'ai mal	e-lio appelle ma fille
Oh là ! Je saigne ! Je me suis blessé	e-lio appelle le SAMU

Table 2: Some examples sentences [in French] identified for inclusion in the scenarios

Before each recording session, the experimenter explained the scenario to the participant. Scenario has following subcomponents: physical location, time of the day, activity (what the person intends to do and what eventually happens), gesture (way of falling), and the spoken sentences. The participant rehearsed the scenario several times before final acquisition of data. On average, the duration of acquisition was 2 hours and 30 minutes per person.

2.3. Equipment and Tools

The studio and the equipment used are shown in Figure 1. Moreover, the DOMUS smart home was equipped with the

social inclusion product e-lio[2].
Regarding audio analysis, we conducted a multi-source capture with two wireless Sennheiser microphones. A eW-300-G2 type ME2 microphone was placed on the suspended ceiling of the studio, and a SKM-300-G2 microphone was placed on a furniture close to the participant. The recording was performed in the technical room by a computer connected to the microphones high-frequency receivers by National Instruments PCI-6220 8-channel card. A loudspeaker in the control room enabled experimenters to hear the progression of the scene in the studio. The National Instruments PCI-6220 card was controlled by the StreamHIS software (Vacher et al., 2011) developed by the GETALP team that enables data acquisition.
Regarding video analysis, we conducted the video capture with two webcam cameras (Sony PSEye: 640x480 60Hz connected by a USB 2.0). Cameras were fixed to the wall, one in front of the Sofa and other at an angle of 45 degrees. Moreover, we acquired depth image with Microsoft Kinect. Synchronization and recording of video streams from all three cameras were performed in the technical room by computer (Intel i3 3.2GHz, 4GB RAM, NVidia GeForce GTS 450 512MB). The tools that enable video acquisition was developed by the team SAARA (based on the OpenCV library). Video processing (background learning and subtraction, body parts segmentation) was performed on a PC cluster, taking advantage of algorithms parallelization on CPU and GPU. Motion tracking and identification of hazardous situations was done on a PC (Intel i7 3.5GHz, 4GB RAM, NVIDIA GTX 660 4GB GeFore).

3. The Acquired Corpus

The CIRDO corpus is divided into three parts, the audio corpus for sound and speech analysis, the video corpus and a third part with the annotations. The corpus is detailed in Sections 3.1. and 3.2..
The audio and video corpus was annotated for the video part, with the Advene software[3], developed at the LIRIS laboratory, and for the sound part, with Transcriber.
The CIRDO corpus is a web-based data library, hosted at LIRIS and accessible for the academic and research purpose. The license to use the CIRDO corpus is granted on a case-by-case basis. We can also grant permission to researchers to extend the corpus given they fulfil the required conditions and sign agreement.

3.1. Audio Corpus

When they played the scenarios, some participants produced sighs, grunts, coughs, cries, groans, panting or throat clearings. As our current studies are in the domain of automatic speech recognition, these sounds were not considered during the annotation process, but this can be done in the framework of future studies. In the same way, speeches mixed with sound produced by the fall were ignored. At the end, each speaker uttered between 10 and 65 short sentences or interjections ("*ah*", "*oh*", "*aïe*", etc.) as shown Table 3.

[2]http://www.technosens.fr
[3]http://liris.cnrs.fr/advene/

Spk.	Nb. of interjections or short sentences		Size
	All	*Call for help*	(s)
Y01	22	14	37.59
Y02	16	15	27.51
Y03	24	21	35.59
Y04	25	15	43.97
Y05	32	21	38.16
Y06	19	15	48.25
Y07	12	12	18.75
Y08	15	12	23.58
Y09	23	21	39.86
Y10	20	19	29.83
Y11	29	27	43.96
Y12	24	21	33.54
Y13	17	14	25.32
S01	65	53	92.07
S02	23	19	31.21
S03	23	21	26.02
S04	24	21	50.33
ALL	413	341	645.54

Table 3: Composition of the audio corpus Cirdo-set

No	Occurrence number of each scenario							Time
	F1	F2	F3	F4	B	TN1	TN2	
Y01	1	2	1	1	1	1	1	8mn 40s
Y02	1	1	1	1	1	1	1	4mn 35s
Y03	1	1	1	3	1	1	2	6mn 30s
Y04	1	1	1	2	1	1	1	5mn 54s
Y05	1	1	1	3	2	1	2	8mn 50s
Y06	1	1	1	1	1	1	1	5mn 07s
Y07	1	1	1	1	1	1	2	5mn 17s
Y08	1	1	2	1	1	1	2	7mn 04s
Y09	3	1	1	1	1	1	2	6mn 48s
Y10	1	1	2	3	1	1	1	5mn 50s
Y11	2	1	2	3	1	1	1	7mn 31s
Y12	1	1	1	2	2	1	2	8mn 01s
Y13	2	2	1	1	1	1	1	5mn 54s
S01	1	1	3	2	3	2	2	9mn 07s
S02	1	2	1	2	1	1	2	6mn 31s
S03	1	1	1	3	1	1	1	6mn 00s
S04	1	1	2	1	2	2	2	7mn 16s
ALL	21	19	23	31	22	19	26	1h 55mn

Table 4: Video Corpus for the two group of participants

Sentences were often close to those identified during the field studies ("*je peux pas me relever* - I can't get up", "*e-lio appelle du secours* - e-lio call for help", etc.), some were different ("*oh bein on est bien là tiens* - oh I am in a sticky situation"). In practice, participants cut some sentences (i.e., inserted a delay between "*e-lio*" and "*appelle ma fille* - call my daughter"), uttered some spontaneous sentences, interjections or non-verbal sounds (i.e., groan)

3.2. Video Corpus

The content of the video corpus is displayed in Table 4

As mentioned above, the video corpus includes various types of fall according to the scenarios defined in the Protocol (refer Section 2.2.). Each scenario includes a background sequence (300-600 frames) which usually is required for human body extraction. Description of the corpus is given in Table 4 wherein F1 to F4 are the fall scenarios, B blocked hip and TN1, TN2 true negatives. The total duration of the recordings is equal to 1 hour and 55 minutes, with a total of 162 video segments.

4. Conclusion Drawing from Feedback from the Participants

Feedback from the participants were analyzed and allowed some conclusions (Body-Bekkadja et al., 2015). A highlighted domestic accident, outlined (for oneself and for others) by CIRDO, leads the elderly individual to relate to the falls in other ways. They suddenly attempt to further regulate and control it, sometimes by risking a profound change in behavior that may not be detected by CIRDO. More specifically, in our various studies (simulated CIRDO use), it was observed that: The elderly developed two opposing conducts in the management of their fall monitored via CIRDO. Some of them chose to boost their falling movement or overplay screams to make sure the device will recognize the danger. They were also the same elderly that, as we mentioned in Study 2, used tangible remote monitoring systems (inset). They doubted the ability of the new ambient system to detect their accident, and thus they risk a more serious injury.

Others, however, in order to hide these incidents from the entourage, try to control the fall: They do not scream, and try to recover at all costs, even though it could exacerbate the deleterious effects of the fall.

In both cases, we see that the falling movement choreography (falling is unintentional by nature) will be intentionally modified by the subjects, whether to be seen to fall, and, hence, be better recognized, or otherwise to be hidden from the technology and its supervision. Thus, the presence of technological artifacts changes the dynamic of the fall, because this behaviour will be addressed, directed toward a target (technology object), and also to other involved individuals. As mentioned by Clot (Clot, 1999), the movements of the falls are not only directed by the conduct of the subject; they are also directed through the use of the technical object, and toward others (the representation they have of the system and the possible recipients of the alert). The fall choreography is influenced by what the individual wants to show to, or hide from, the device. Therefore, the risk of non-detecting these movements is possible, because they sometimes tend to veer too far from programmed scripts.

Another more symbolical consequence involves the status and the legitimacy that CIRDO will give the fall. While, in the past, the word of the victim, or of a third party, used to be enough to prove the occurrence of a fall, these days, surveillance technologies are used to validate/approve the occurrence of an incident. We could even add that, they also assess whether this fall is acceptable, and conform to pre-defined scripts. In other words, the non-detection (non-recognition) of a fall by CIRDO may mean that: (1) The individual has poorly conducted the choreography of the falling; or (2) the movement cannot be categorized as a risk movement. This can therefore lead to a denial of risk and recognition of the elderly individual's status of "victim", thereby throwing suspicion on their statements.

As a result, CIRDO may operate a transfer of the risk recognition: It is the technology that gives official status to a fall in a domestic incident. It gives credibility and legitimacy to it. Thus, technological reliability wins against the word of the elderly; and then the latter can be discredited, for example, if the user talks about incidents not recognized by technology: "If the system is not triggered, then the fall did not occur." In both cases, setting CIRDO service in the domestic social system can either redefine the falls choreography, or redefine their status. All these reasons may be caused by malfunction or rejection.

Several recommendations can be addressed to the designers. In addition to reassuring, involving, and properly training to use the new system (Hwang and Thorn, 1999)(Kujala, 2003), by including demonstrations and updates in real situations adapted to their habits and lifestyle practices (to demonstrate the system works efficiently in detecting falls, and avoiding any worsening of movements), designers should create programs and falls-detection algorithms more flexible, to cover a larger spectrum of falls choreography, and not rely solely on fixed scripts in risk behavior. In the same way, effective articulation between Audio and Video detections should be ensured to allow better falls-recognition and validation.

As part of a participatory design through use (He and King, 2008), it would be interesting to adjust from feedbacks in the field, the location of the sensors and the level of falls scripts detection, by taking into account conduct-adjustments and daily risk evolutions from the moment the technological artifact is introduced. This requires a situated analysis of the actual device's uses, in order to re-design it from its usage (Norman and Draper, 1986). This reinforced monitoring system will ask the elderly to make an oral (confirmation) emergency call to an outside third party. This alert is then triggered only in case of a positive response, or a lack of response from the individual, enabling them to properly control the system (at to avoid false alarms). Finally, this new device for monitoring the activity will be meaningful only if it serves and supports its users' quality of life. Therefore, one must be careful to involve and assist the various participants during all phases of CIRDO design, in an inclusive approach (as in the Living lab approach of (Pino et al., 2015), especially to anticipate reconfigurations at work in the social framework, as we shall now see.

5. Call for Help Recognition from Audio Analysis

The audio corpus was used for studies related to call for help recognition in Distant Speech conditions thanks to online speech analysis using SGMM acoustic modelling (Vacher et al., 2015a) (Vacher et al., 2015b). Non-speech analysis might bring information of high interest but it was not considered in our study because of the lack of training data. Currently non-speech audio analysis is a relatively unexplored field due not only to the lack of data but also to the unexpected sound classes that can be recorded at home in unconstrained conditions (Vacher et al., 2011). Moreover, non-speech sounds are made of non-verbal sounds (i.e. groans, panting, throat clearings, etc.) and of daily living sounds (i.e. falling objects, door slap, etc.) which represent different semantic information and are difficult to differentiate. For these reasons, only speech analysis was considered.

5.1. Acoustic modelling

The Kaldi speech recognition tool-kit (Povey et al., 2011b) was chosen as ASR system. Kaldi is an open-source state-of-the-art Automatic Speech Recognition (ASR) system with a high number of tools and a strong support from the community. In the experiments, the acoustic models were context-dependent classical three-state left-right HMMs. Acoustic features were based on Mel-frequency cepstral coefficients, 13 MFCC-features coefficients were first extracted and then expanded with delta and double delta features and energy (40 features). Acoustic models were composed of 11,000 context-dependent states and 150,000 Gaussians. The state tying is performed using a decision tree based on a tree-clustering of the phones. In addition, off-line fMLLR linear transformation acoustic adaptation was performed.

The acoustic models were trained on 500 hours of transcribed French speech composed of the ESTER 1&2 (broadcast news and conversational speech recorded on the radio) and REPERE (TV news and talk-shows) challenges as well as from 7 hours of transcribed French speech of the SWEET-HOME corpus (Vacher et al., 2014) which consists of records of 60 speakers interacting within a smart home and from 28 minutes of the Voix-détresse corpus (Aman, 2014) which is made of records of speakers eliciting a distress emotion.

5.1.1. Subspace GMM Acoustic Modelling

The GMM and Subspace GMM (SGMM) both model emission probability of each HMM state with a Gaussian mixture model, but in the SGMM approach, the Gaussian means and the mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections.

The SGMM model (Povey et al., 2011a) is described in the following equations:

$$\begin{cases} p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}. \end{cases} \quad (1)$$

where $\mathbf{x}$ denotes the feature vector, $j \in \{1..J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$ i.e; the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ are common across all states. These parameters can be shared and estimated over multiple record conditions.

A generic mixture of I Gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialization of the SGMM.

Our experiments aims at obtaining SGMM shared parameters using both SWEET-HOME data (7h), Voix-détresse (28mn) and clean data (ESTER+REPERE 500h). Regarding the GMM part, the three training data set are just merged in a single one. (Povey et al., 2011a) showed that the model is also effective with large amount of training data. Therefore, three UBMs were trained respectively on SWEET-HOME data, Voix-détresse and clean data. These tree UBMs contained 1K gaussians and were merged into a single one mixed down to 1K gaussian (closest Gaussians pairs were merged (Zouari and Chollet, 2006)). The aim is to bias specifically the acoustic model towards distant speech home and expressive speech conditions.

5.2. Recognition of distress calls

The recognition of distress calls consists in computing the phonetic distance of an hypothesis to a list of predefined distress calls. Each ASR hypothesis H_i is phonetized, every predefined voice command T_j is aligned to H_i using Levenshtein distance. The deletion, insertion and substitution costs were computed empirically while the cumulative distance $\gamma(i,j)$ between H_j and T_i is given by Equation 2.

$$\gamma(i,j) = d(T_i, H_j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (2)$$

The decision to select or not a detected sentence is then taken according a detection threshold on the aligned symbol score (phonems) of each identified call. This approach takes into account some recognition errors like word endings or light variations. Moreover, in a lot of cases, a miss-decoded word is phonetically close to the good one (due to the close pronunciation). From this the CER (Call Error Rate i.e., distress call error rate) is defined as:

$$\text{CER} = \frac{\text{Number of missed calls}}{\text{Number of calls}} \quad (3)$$

This measure was chosen because of the content of the corpus Cirdo-set used in this study. Indeed, this corpus is made of sentences and interjections. All sentences are calls for help, without any other kind of sentences (e.g., colloquial sentences), and therefore it is not possible to determine a false alarm rate in this framework.

5.3. Off line experiments

The methods presented in previous sections were run on the Cirdo-set audio corpus presented in Section 3.1.

The SGMM model presented in Section 5.1. was used as acoutic model. The *generic language model* (LM) was estimated from French newswire collected in the Gigaword corpus. It was 1- gram with 13,304 words. Moreover, to reduce the linguistic variability, a 3-gram domain language model, the *specialized language model* was learnt from the sentences used during the corpus collection described in Section 2.2., with 99 1-gram, 225 2-gram and 273 3-gram models. Finally, the language model was a 3-gram-type which resulted from the combination of the *generic LM* (with a 10% weight) and the *specialized LM* (with 90% weight). This combination has been shown as leading to the best WER for domain specific application (Lecouteux et al., 2011). 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 avoid the recognition of sentences leading to "false-positive" detection.

Spk.	WER (%)		CER (%)	Spk.	WER (%)		CER (%)
	All	*Call for help*			*All*	*Call for help*	
Y01	45.0	39.1	27.8	Y07	21.3	17.0	16.7
Y02	41.4	44.4	40.0	S04	30.8	25.0	25.0
S01	51.9	49.6	34.0	Y08	45.9	43.6	23.8
Y03	19.1	15.4	14.3	Y09	67.0	54.8	50.0
S02	39.2	34.3	26.3	Y10	21.5	19.5	5.3
S03	21.2	20.3	28.6	Y11	14.9	11.76	7.4
Y04	61.8	50.8	20.0	Y12	21.4	22.4	19.0
Y05	49.4	41.2	33.3	Y13	57.7	44.9	71.4
Y06	24.5	22.4	14.3	All	39.3	34.0	26.8

Table 5: Word and Call Error Rate for each participant

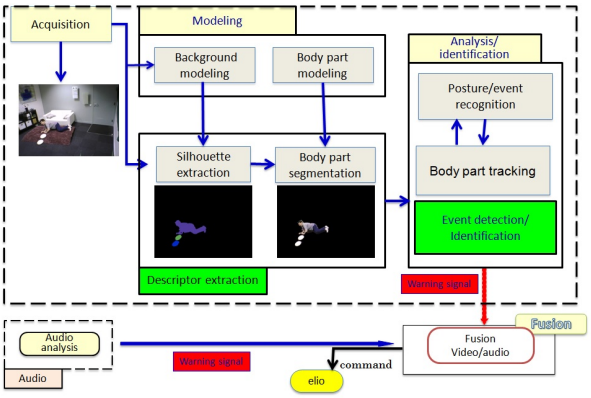

Figure 3: Event detection, video pipeline.

Results on manually annotated data are given Table 5. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific distress calls as well as the Call Error Rate (CER: c.f. equation 3). Considering distress calls only, the average WER is 34.0% whereas it a 39.3% when all interjections and sentences are taken into account.

On average, CER is equal to 26.8% with an important disparity between the speakers.

Unfortunately and as mentioned above, the used corpus does not allow to determine a False Alarm Rate. Previous studies based on the AD80 corpus showed recall, precision and F-measure equal to 88.4%, 86.9% and 87.2% (Aman et al., 2013). Nevertheless, this corpus was recorded in very different conditions, text reading in a studio, in contrary of those of Cirdo-set.

6. Fall Detection from Video Analysis

The video analysis framework for fall events labeling is based on the silhouette extraction. The silhouette is extracted by removal of background pixels within the video scene using. The background subtraction is performed by Mixture of Gaussian based approach, and eigenbackground (PCA) approach (Deeb et al., 2012), (Priyank Shah, 2014). The extracted shape is then used to obtain discriminative features to detect anomaly in person's movement in the scene. The overview of proposed framework is shown in Figure 3.

6.1. Foreground extraction and modeling of body parts

After silhouette extraction the human body is segmented into colored connected components. Colored connected components regarded as regions are extracted by applying region growing technique. Each component with similar color are fitted into blob (Hsieh et al., 2010). Blob is represented by spatial color Gaussian mixture model, assuming that spatial color components are de-correlated. The probability of an observation X_t of that pixel at time t to belong to the background is given by equation 4.

$$P(\mathrm{X}_t) = \sum_{k=1}^{L} \omega_{k,t}.\eta(\mathrm{X}_t, \mu_{k,t}, \Sigma_{k,t}) \quad (4)$$

where L is the number of gaussian, η is Gaussian probability density function, $\Sigma_{k,t}$ is an estimates of weight and $\omega_{k,t}$ is covariance matrix of the k-th Gaussian in the mixture at time t. In the last, an energy based function is formulated over the unknown labels of every pixel in the form of a first order Markov random field (MRF) energy function:

$$E(f) = \sum_{p \in P} D_p(f_p) + \lambda \sum_{p,q \in N_e} V_{p,q}(f_p, f_q) \quad (5)$$

Here, N_e is neighboring pixels and the data energy $\Sigma_{p \in P} D_p(f_p)$ evaluates the likelihood of each pixel to take a label. Finally, energy function is minimized with a graph cut algorithm via a swap approach (Miguel Angel Bautista, 2015). Background model is updated selectively with an online EM algorithm (Moon, 1996).

6.2. Tracking of body parts

Our method begins with an iterative process and considers that person is in standing position. Body is decomposed in three parts i.e. head, torso and lower part. Then, dynamic processing technique is used to obtain different level of body details (i.e. number of blobs), if required. At first, a blob (k-th blob) is built for each connected component with a feature vector containing: color, center's coordinates C_k, a predicted change of color P_k and velocity V_k. Set of current colored connected components are extracted from the smoothed foreground by region growing. By considering tracking of blobs as a matching process, we introduce a novel cost function to obtain distance between two blobs. The cost function is given by:

$$Cost(j,k) = \alpha_1(\|P_k - P_j^t\|/V_k) + \alpha_2(\|C_j - C_k^M\|/\bigtriangledown C_k) \quad (6)$$

where C_j^t and P_j^t respectively, the color and the center's coordinates of the j-th connected component extracted at time t.

6.3. Pose recognition and labeling of events

The labeling of events is based on the scenario define above. To identify poses, the extracted silhouette from the body parts was used. To recognize and label different poses from the video data, an histogram based approach was used as they runs in real time and improves its accuracy with the passage of run time The recognition wss based

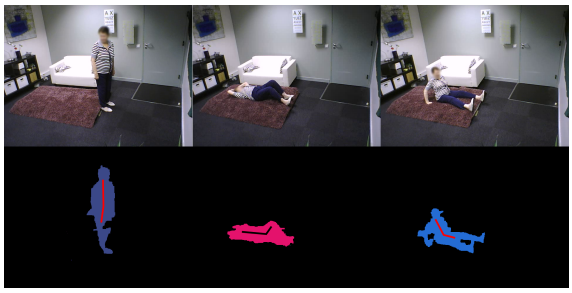

Figure 4: Event detection, results

on comparison of two histograms i.e. key pose frame histogram and histogram of frame in hand (Barnachon et al., 2012). The histograms were then compared with the Bhattacharyya distance and warped by a dynamic time warping process to achieve their optimal alignment (Barnachon et al., 2014). Bhattacharyya distance is calculated using:

$$d(H_1, H_2) = \sqrt{1 - \Sigma_I \frac{\sqrt{H_1(I).H_2(I)}}{\sqrt{\Sigma_I H_1(I).\Sigma_I H_2(I)}}} \quad (7)$$

Where H_1 and H_2 are integral histograms.
In this study we focused on the analysis of posture to detect event of distress. Our framework for fall detection could further be enhanced by considering other factors as well, i.e. facial expression analysis. Analysis of facial expressions to detect pain can be very advantageous in case of prolonged immobility of elderly person (Khan et al., 2013).

7. Conclusion

This paper investigated smart home technology and recorded comprehensive audio/video corpus of domestic falls. For elderly, the fall is one of the most feared and recurring problems. Surveillance technologies tries to provide solution to this issue by alerting contact person in case of fall. There is lot of work which focuses on this problem but to the best of our knowledge there is no common database for comparing results. To address this issue we have made available a database composed of audio/video records of falls and prolonged immobility. In this paper, studies using this database for the automatic identification of these events were also presented.
The CIRDO corpus was recorded in a smart environment reproducing a typical living room containing an e-lio communication device equipped with microphone and camera. The experiment was guided by an ethnographic study which detailed the various events related to distress situations faced by elderly people. Various scenarios were established from this study that describes pattern of postures in case of fall so that acted falls were as close to real situation as possible. The database can thus be useful to researchers of the community studying video or audio emergency situations as well as to model the relationship between audio/video events and learn fall model using machine learning.

Acknowledgements

This work was supported by the French funding agencies ANR and CNSA through CIRDOproject (ANR-2010-TECS-012). The authors would like to thanks the persons who agreed to participate in the survey or in the recordings.

Bibliographical References

Aman, F., Vacher, M., Rossato, S., and Portet, F. (2013). Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences. In *The 7th Int. Conf. on Speech Technology and Human-Computer Dialogue, SpeD 2013*, pages 177–184.

Aman, F. (2014). *Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile*. Ph.D. thesis, Université de Grenoble, Ecole doctorale MSTII.

Barnachon, M., Bouakaz, S., Boufama, B., and Guillou, E. (2012). Human actions recognition from streamed motion capture. In *Pattern Recognition (ICPR), Int. Conf. on*, pages 3807–3810. IEEE.

Barnachon, M., Bouakaz, S., Boufama, B., and Guillou, E. (2014). Ongoing Human Action Recognition with Motion Capture. *Pattern Recognition*, 47(1):238–247.

Bloch, F., Gautier, V., Noury, N., Lundy, J., Poujaud, J., Claessens, Y., and Rigaud, A. (2011). Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects. *Annals of Physical and Rehabilitation Medicine*, 54:391–398.

Bobillier-Chaumon, M.-E., Cuvillier, B., Bouakaz, S., and Vacher, M. (2012). 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*, pages 121–122, Dijon, France.

Body-Bekkadja, S., Bobillier-Chaumon, M.-E., Cuvillier, B., and Cros, F. (2015). Understanding the Socio-Domestic Activity: A Challenge for the Ambient Technologies Acceptance in the Case of Homecare Assistance. In *HCI International*, volume Part II of *LNCS 9194*, pages 399–411.

Bouakaz, S., Vacher, M., Bobillier-Chaumon, M.-E., FrédéricAman, Bekkadja, Portet, Guillou, E., Rossato, S., Desserée, E., Traineau, P., Vimon, J.-P., and Chevalier, T. (2014). CIRDO: Smart companion for helping elderly to live at home for longer. *Innovation and Research in BioMedical engineering*, 35(2):101–108.

Clot, Y. (1999). *La fonction psychologique du travail*. PUF, Paris, France.

Deeb, R., Desseree, E., and Bouakaz, S. (2012). Real-time two-level foreground detection and person-silhouette extraction enhanced by body-parts tracking. In *Proc. SPIE*, volume 83010R, pages 1–8.

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

He, J. and King, W. (2008). The role of user participation in information systems development: Implications from a meta-analysis. *Journal of Management Information Systems*, 25(1):301–331.

Hsieh, C., Chuang, S., Chen, S., Chen, C., and Fan, K. (2010). Segmentation of human body parts using deformable triangulation. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 40(3):596–610.

Hwang, M. and Thorn, R. (1999). The effect of user engagement on system success: A meta-analytical integration of research findings. *Information and Management*, 35(4):229–236.

Igual, R., Medrano, C., and Plaza, I. (2013). Challenges, issues and trends in fall detection systems. *BioMedical Engineering OnLine*, 12(1):1–24.

Kangas, M., Vikman, I., Nyberg, L., Korpelainen, R., Lindblom, J., and Jämsä, T. (2012). Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects. *Gait & Posture*, 35(3):500 – 505.

Khan, R. A., Meyer, A., Konik, H., and Bouakaz, S. (2013). Pain detection through shape and appearance features. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6, July.

Klenk, J., Becker, C., Lieken, F., Nicolai, S., Maetzler, W., Alt, W., Zijlstra, W., Hausdorff, J., van Lummel, R., Chiari, L., and Lindemann, U. (2011). Comparison of acceleration signals of simulated and real-world backward falls. *Medical Engineering & Physics*, 33(3):368 – 373.

Kujala, S. (2003). User involvement: A review of the benefits and challenges. *Behaviour and Information Technology*, 22(1):1–16.

Lecouteux, B., Vacher, M., and Portet, F. (2011). Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In *Interspeech 2011*, pages 1–4, Florence, Italy.

Miguel Angel Bautista, Sergio Escalera, D. S. (2015). Hupba8k: Dataset and ecoc-graph-cut based segmentation of human limbs. *Neurocomputing*, 150(A):173–188.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60.

Norman, D. A. and Draper, S. W. (1986). *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc.

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2006). Human computing and machine understanding of human behavior: survey. In *ACM Int. Conf. on Multimodal Interfaces*.

Pino, M., Moget, C., Benveniste, S., Picard, R., and Rigaud, A.-S. (2015). Innovative technology-based healthcare and support services for older adults: How and why industrial initiatives convert to the living lab approach. In *HCI International*, volume Part II of *LNCS 9194*, pages 158–169. Springer International Publishing.

Popescu, M., Li, Y., Skubic, M., and Rantz, M. (2008). 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*, pages 4628–4631, 20–25 Aug.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., Rose, R. C., Schwarz, P., and Thomas, S. (2011a). The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, 25(2):404 – 439.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011b). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Priyank Shah, H. M. (2014). Comperehensive study and comparative analysis of different types of background subtraction algorithms. *International Journal of Image,Graphics and Signal Processing*, 6(8):47–52.

Saim, R. (2014). Accidental falls amongst the elderly: Health impact and effective intervention strategies. Master's thesis.

Vacher, M., Portet, F., Fleury, A., and Noury, N. (2011). Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges. *International journal of E-Health and medical communications*, 2(1):35–54, March.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

Vacher, M., Aman, F., Rossato, S., and Portet, F. (2015a). Development of Automatic Speech Recognition Techniques for Elderly Home Support: Applications and Challenges. In J. Zou et al., editors, *HCI International*, volume Part II of *LNCS 9194*, pages 341–353, Los Angeles, CA, United States, August. Springer International Publishing Switzerland.

Vacher, M., Lecouteux, B., Aman, F., Rossato, S., and Portet, F. (2015b). Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home. In *6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 1–7, Dresden, Germany, September. SIG-SLPAT.

Who. (2013). World population ageing: 1950-2050. Technical report, Executive report, United Nations, Department of Economic and Social Affairs, Population Division.

Young, V. and Mihailidis, A. (2010). An automated, speech-based emergency response system for the older adult. *Gerontechnology*, 9(2):261.

Young, V. and Mihailidis, A. (2013). The CARES corpus: a database of older adult actor simulated emergency dialogue for developing a personal emergency response system. *I. J. Speech Technology*, 16(1):55–73.

Zouari, L. and Chollet, G. (2006). Efficient gaussian mixture for speech recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 294–297.