A multi-sensor approach for People Fall Detection in home environment

A. Leone, G. Diraco, C. Distante, P. Siciliano, M. Malfatti, L. Gonzo, M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, et al.

A multi-sensor approach for People Fall Detection in home environment

A. Leone[1], G. Diraco[1], C. Distante[1], P. Siciliano[1], M. Malfatti[2], L. Gonzo[2], M. Grassi[3], A. Lombardi[3], G. Rescio[3], P. Malcovati[3], V. Libal[4], J. Huang[4], and G. Potamianos[4]

[1] Institute for Microelectronic and Microsystems CNR-IMM 73100, Lecce, Italy
[alessandro.leone, giovanni.diraco, cosimo.distante]@le.imm.cnr.it
[2] Integrated Optical Sensor Group FBK-irst 38050, Povo (TN), Italy
[3] Department of Electrical Engineering University of Pavia 27100, Pavia, Italy
[4] T.J. Watson Research Center IBM 10598, Yorktown Heights (NY), USA

Abstract. This paper presents a hardware and software framework for reliable fall detection in the home environment, with particular focus on the protection and assistance to the elderly. The integrated prototype includes three different sensors: a 3D Time-Of-Flight range camera, a wearable MEMS accelerometer and a microphone. These devices are connected with custom interface circuits to a central PC that collects and processes the information with a multi-threading approach. For each of the three sensors, an optimized algorithm for fall-detection has been developed and benchmarked on a collected mulitimodal database. This work is expected to lead to a multi-sensory approach employing appropriate fusion techniques aiming to improve system precision and recall.

1 Introduction

Over the past few years, research has been increasingly focusing on building systems for observing humans and understanding their appearance and activities. Furthermore, home assistance and protection became an important topic in sensors systems. In particular in the work reported in this paper, emerging sensing technologies [1], [2] are exploited in order to detect possible falls [3] and disease of older people in their own home environment, delivering an alarm flag to emergency operators or relatives. On one side solutions based on this kind of approach are a social advantage first of all regarding assisted people themselves since they are able to continue living in their own familiar environment, while also from care-services delivery functionality point of view it may be verified to be strongly convenient. In fact European population ageing 65 years or more, which may be in need of assistance is getting wider and wider i.e. of the order of 40 millions in year 2000 and expected to be around 55 millions before year 2025. This trend asks care-holders institutions to employ more efficient and optimized methods in order to be able to grant the required service. For this reason various projects and consortia have been funded and created under European Community coordination, including Netcarity [4], the one in which the work

described in the following manuscript has been developed. Netcarity proposes a new integrated paradigm to support independence and inclusion in ageing people living alone at home. The project fosters the development of a light technological infrastructure to be integrated in the homes of old people at reduced costs, that allows both the assurance of basic support for everyday activities and detection of critical health situations, as well as the social and psychological engagement required to maintain emotional well being in the elder, enhancing dignity and quality of life. From the bare sensing technology development point of view the aim of the project is the design of a network multisensory system thought for older people assistance in home environment in terms of health care, safety and security. In particular in this paper, as mentioned, the focus will be on the development of a high-recall room-ranging people position/fall detector. It includes three different sensors: a 3D Time-of-Flight range camera, a wireless wearable MEMS accelerometer and a professional microphone. These devices are connected with ad-hoc interface circuits to a central host embedded PC (e_PC) that receives and processes the information with a multi-threading-approach. All the information gathered from the sensors is only temporary stored on the e_PC hard-drive for the processing needed time and only help request and its motivation will be transmitted to care-holders outside the house in order to completely fulfill assisted person privacy.

2 Position detector framework overview

The hardware framework is reported in Fig. 1 and the block diagram of the complete system is depicted in Fig. 2. The information provided by the employed 3D camera allows us to describe the environment quantitatively, in terms of appearance and depth images at QCIF resolution. In addition, a wearable three-axis accelerometer [5], by means of a ZigBee wireless module, delivers acceleration components to the e_PC, which recognizes specific patterns related to falls. Finally, acoustic scene analysis is performed using an off-the-shelf Shure microphone. As already mentioned, our goal is to improve the state-of-the-art by exploiting a multi-sensory approach for fall detection, adopted to minimize false alarms. For this purpose, the data delivered by each sensor is first processed by separate algorithms, as described in this paper. Eventually, data fusion over a given time window will be performed subsequently by the e_PC. A brief outline of the employed devices is given in the following.

2.1 3-D Camera

In the last years several active range sensors having real time performances have been presented. The ability to describe scenes in three dimensions opens new scenarios, providing new opportunities in different applications, including visual monitoring (object detection, tracking, recognition, image understanding, etc.) and security contexts. Among all active range sensors, Time-Of-Flight (TOF) range cameras present several advantages in the use (i.e. small dimensions, low

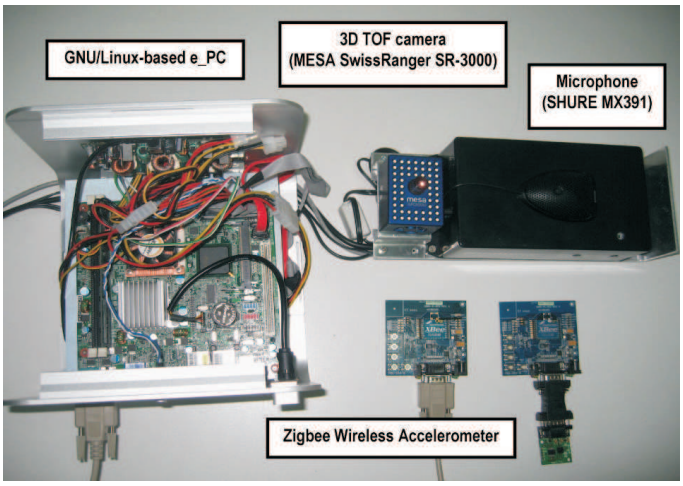


Fig. 1. The proposed position detector framework architecture based on commercial sensors.

power consumption), integrating distance measurement as well imaging aspects (RIM - Range Imaging). The vision sensor used in this work is a Time-Of-Flight camera MESA SwissRanger 3000 [6]. The wall-mounted camera is able to measure both the appearance information (gray levels image) of the scene and the depth (in meters) for each pixel. The information coming from the camera is processed to extract the blobs (moving regions), detect moving people and track them in the scene. The employed commercial 3D camera provides two kind of images at QCIF spatial resolution (a depth map and a grey levels image) at least 15 fps (up to 25fps, variable with setting parameters). The camera works with an integrated, modulated near infrared light source (the illuminator is composed by an array of 55 near infrared LEDs). The emitted light is reflected by the objects in the scene and sensed by a pixel array realized in a specialized mixed CCD and CMOS process: the target depth is estimated by measuring the phase shift of the signal round-trip from the device to the target and back. Moreover, attention must be dedicated to the Field-Of-View (FOV) of the device: normally active range sensor exhibits narrow FOV (47.5×39.6 $V \times H$ degs) so a pan-tilt solution could be used for monitoring large areas. A narrow-band infrared filter is used so that depth map and intensity image are not affected by environmental illumination conditions (the camera is suitable in night vision applications and allow to use computer vision algorithms much less complex). By using the default parameters, the SR camera is able to define a depth map with a good approximation (greater than 99%) when the target object falls in the non-ambiguity range (up to 7,5 meters when the modulation frequency is 20Mhz). When the camera-target distance overcomes the non-ambiguity range, aliasing effects appear (i.e. object at 8 meters is "seen" as at 0,5 meters), demoting only the depth estimation. Another important parameter is the Integration Time, that

can be tuned in a proper way to limit saturation and noise effects. In particular, if the Integration Time is short the results are very noisy, if it is long the results are smoothed and overflow starts to contaminate the results with objects close to the camera. At the moment, the raw data provided by the camera do not reach levels of accuracy generally required in industrial applications. Therefore, a calibration of the sensor is strongly required to recover more accurate metric information. Outputs provided by camera via USB 2.0 connection include both raw data and on-board FPGA processed data for noise reduction, correction and 3-D coordinates evaluation, so that for each frame the vision-based fall detector computes almost 396Kbytes of information. The above discussion is also suitable for other active RIM sensors using similar technologies, as they exhibit the same problems (non-ambiguity range, etc.). Table 1 summarizes the main characteristics of the optical device.

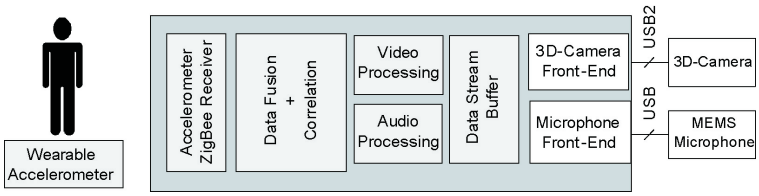


Fig. 2. Block diagram of the proposed hardware-software framework.

Table 1. 3-D CAMERA MAIN CHARACTERISTICS

Specifications	
Pixel Array Size	176 × 144 (QCIF)
Field of View	47.5 × 39.6 degrees
Interface	USB 2.0
Illumination Power	1 Watt @850nm
Power Supply	12V
Power Consumption	12 W, typical
Operating Temperature	-10C to +50C
Output Data (per pixel)	x, y, z coordinates i (intensity)
Range and Resolution	
Modulation Frequency	20MHz, standard
Non-ambiguous range	7.5 meters
Distance Resolution	1% of range, typical
Frame Rate	25 fps, typical

2.2 Wireless accelerometer

The 3-axis integrated accelerometer circuit core, in this preliminary version of the module, is the Freescale Semiconductor MMA7260Q chip [7]. The device may operate between 2.2V and 3.6V, allowing the use of batteries without additional voltage regulators. The acceleration bandwidth which may be processed by the micro-machined device is about 150 Hz, which has been demonstrated to be sufficient to detect falls in preliminary testing. Also full-scale values available for the chosen device, which lie between $\pm 1g$ to $\pm 6g$ are satisfactory for this system. In the preliminary data collection, the full-scale has been set to $\pm 2g$, in order to provide maximum sensitivity, while avoiding saturation. Analog acceleration information is processed by the 10-bit A/D converter of a PIC micro-controller and delivered to the wireless module as a UART serial data stream (in streaming mode). The data size of a single complete 3D-acceleration packet is 11 Bytes and, therefore, for the full sensor bandwidth a serial bit-rate of 38.4 kbps is sufficient to transmit the data-set without loss. The employed ZigBee wireless module fulfills this requirement, being able to transmit data from 9600 bps to 250 kbps. With the available acceleration full-scale values, the 10-bit acceleration information leads to an output sensitivity from $0.003g$ for minimum FS range to $0.012g$ for maximum FS range. This preliminary module, as mentioned, has been basically developed for studying the algorithms and has not yet been optimized in terms of power efficiency. In the final device, fall detection algorithms will be loaded on an onboard FPGA, and the acceleration streaming mode will be replaced by a fall-flag transmission mode, turning on the transmitter only in case of an alarm.

2.3 Microphone

The audio data from the microphone are analyzed for detecting sound patterns typical of a fall as well as requests for help. A hidden Markov model based approach is used for this purpose. In order to listen to the environment to get fall noise patterns or requests for help a commercial Shure microphone is employed, directly connected to the input of the 16-bit audio card of the embedded PC. Shure Microflex MX391 Series microphones [8] are small, surface mounted electret condenser microphones designed for mounting on conference tables, stage floors, and lecterns. Their high sensitivity and wide frequency range make it especially suitable for picking up speech and vocals in sound reinforcement and recording applications. Flat frequency response across the vocal range for uncolored sound and interchangeable cardioid, supercardioid, and omnidirectional cartridges in order to provide optimal choice for each application are the most interesting features of the employed product for the aim of the framework.

3 Algorithms and experimental data

The following sub-sections describe the architecture of the single sub-system algorithms, whose outputs represent suitable input variables for the envisaged

sensor fusion process. Each sub-system presents a set of different possible events recognized by the processing algorithms, together with a probability distribution which indicates the confidence level of the recognized event (fall). For every given sensor, an example of data acquisition and interpretation is also given. The data acquisition campaign has been performed with 13 different actors that produced more than 450 events including approximately 210 falls. A total of 32 sessions are annotated with different labels for each type of fall. Among them, 20 sessions are used for training, 5 sessions are used as a held-out set, and 7 sessions (containing data from 3 separate actors) are used for algorithm testing. Detection performance is evaluated in terms of recall - i.e., the number of correct system output events divided by the number of system output events - and efficiency - namely, the number of correctly detected reference events divided by the number of reference events.

3.1 3D imaging analysis

The 3D vision-based fall detector has been implemented in C++ on a Linux-based architecture. The main steps of the algorithmic framework are discussed in the following. In the first activity, a background model is estimated and a depth image of the scene is provided where no people or moving objects are presented. The background modeling is realized according to a statistical technique known as Mixtures of Gaussians (MoGs) [9] in which 3 Gaussians are used. The detection of a moving person is obtained through a Bayesian segmentation. The use of depth information provides important improvements in the segmentation, since it is not affected by typical problems of the traditional passive vision such as shadows, mimetic appearances and severe illumination conditions (Fig. 3).

The tracking/estimating of human body motion present many hurdles due to the complexity and variability of the appearance of the human body, the nonlinear nature of human motion, and a lack of sufficient image cues about 3D body pose, including self-occlusion. It is important to note that tracking can increase the robustness, especially when occlusions occur or when objects temporally disappear. The standard approach for tracking is to use a Kalman Filter for every object. This, however, requires the use of a high complexity management system to deal with the multiple hypotheses necessary to track objects. For the previous reason, in this context we adopt a stochastic approach which is based on the ConDensation algorithm (Conditional Density Propagation over time [10]) that is able to perform tracking with multiple hypotheses in range images. A probability density function describing the likely state of the objects is propagated over time using a dynamic model. The measurements influence the probability function and allow the incorporation of new objects into the tracking scheme. The motivation for choosing this new tracking approach is its ability to easily track multiple hypotheses and its simplicity. In addition, a constant computing time can be ensured which is very helpful for real-time applications. In the first prototype of the fall-detector the position of the centroid of the extracted blob in the image plane from depth information is predicted frame-by-frame according to the previous discusion. In particular, a simple state vector is

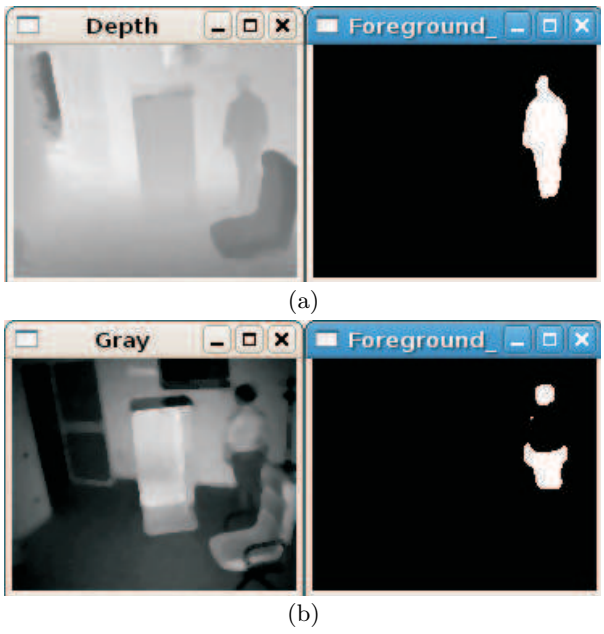


Fig. 3. The binary Bayesian segmentation is better when depth information (a) is used. The appearance properties of people/background cause an inaccurate segmentation (b) when only intensity information is used.

used by merging the following information: (x, y, z)-coordinates of the centroid and the corresponding speeds along the (x, y, z)-axis. The tracker is realized by filtering (thresholding) the Euclidean distance between the predicted location of the centroid and its measured version in the adjacent time step. A preliminary study shows that the error in the prediction step is always lower than 1 pixel (Fig. 4). Once the foreground is extracted and the people is tracked in the scene, the vision-based sub-system analyzes the segmented blob to see if the condition of fall event is met: a crucial step is the extraction of reliable features to detect critical behaviors (falls). For the proposed architecture, the distance of the centroid of the segmented blob from the ground-floor is evaluated through a system coordinates changing transformation. Since the 3-D points from the TOF sensor are represented in a coordinate system centered in camera optic center (in which camera coordinate system is $O(X, Y, Z)$ as shown in Fig. 5), a coordinate change from $O(X, Y, Z)$ to $O'(X', Y', Z')$ is needed to estimate the height of the human centroid (known as h in Fig. 6). The coordinate transformation can be thought as a rotation around X axis, when the rotation around Z axis is negligible.

Let $h = h' + H$ the centroid height, where H is the distance of the camera from the ground floor and $h' = (P_Z)_{O'}$ is the z-coordinate of the P point in the $O'(X', Y', Z')$ system coordinate, according to the previous Fig. 6. The system

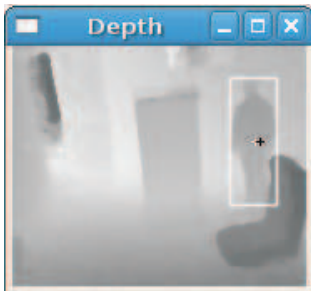
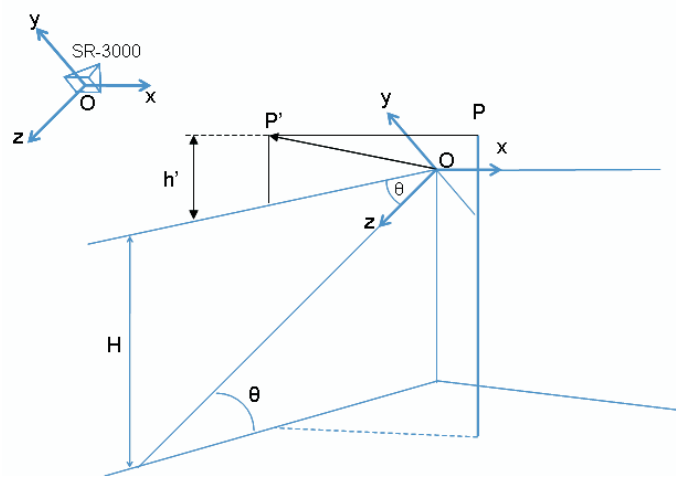


Fig. 4. The black cross is referred to the estimated (predicted) position of the centroid in the image plane when the ConDensation algorithm is applied. The white cross shows the measured location of the centroid related to the extracted bounding box (white rectangle). The error in the prediction is normally lower than 1 pixel.



h

Fig. 5. Camera pose geometry of the acquisition system.

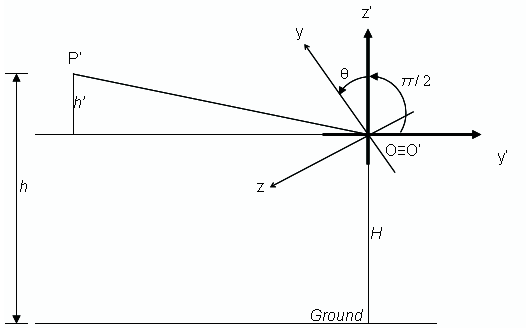


Fig. 6. Coordinate rotation around the x-axis.

transformation matrix is defined as:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta + \frac{\pi}{2} & -\sin\theta + \frac{\pi}{2} \\ 0 & \sin\theta + \frac{\pi}{2} & \cos\theta + \frac{\pi}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\sin\theta & -\cos\theta \\ 0 & \cos\theta & -\sin\theta \end{bmatrix} \tag{1}$$

where $\theta + \pi/2$ is the angle of rotation. The P point is transformed according to the following relation:

$$(P)_{O'} = R \cdot \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = \begin{bmatrix} P_x \\ -P_z\cos\theta - P_y\sin\theta \\ P_y\cos\theta - P_z\sin\theta \end{bmatrix} \tag{2}$$

By substituting, the centroid height with respect the ground floor is defined as:

$$h = H + P_y\cos\theta - P_z\sin\theta \tag{3}$$

A fall event is detected by thresholding the height h of the centroid (in meters) from the floor. In particular, the fall event is characterized 1) by a centroid distance lower than a prefixed value (0.4 meters are a good choice in our experimental setup to avoid false alarms) and 2) an unchangeable situation (negligible movements in the segmented image) for at least 15 consecutive frames (about 1.5 seconds). Fig. 7 shows the typical pattern of the centroid distance when fall events occur (red circles): in these situations the distance is lower than the threshold and the ad-hoc software detects the critical events. In order to evaluate the performance of the 3D vision-based fall detector, a 2 classes (fall, non-fall) clustering approach has been defined and the confusion matrix has been studied to evaluate performance metrics (table 2). The performance is affected especially by several misclassifications due to total occlusion situations (i.e. the person falls behind a big object). Again, the performance demotes when a poor segmentation occurs due to an unstable background model (i.e. the fall occurs during the transient period of the background modeling).

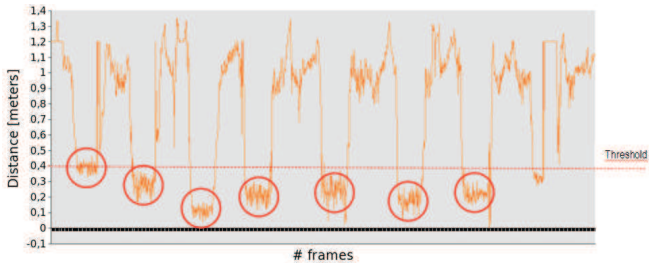


Fig. 7. Pattern of the centroid height.

Table 2. 3D VISION ALARM EFFICIENCY AND RELIABILITY

Threshold for centroid height	**Precision**	**Recall**
0.3 meters	51.1%	99.2%
0.4 meters	80.0%	97.3%
0.5 meters	81.3%	89.3%

3.2 Accelerometer analysis

Analyzing the output of each of the accelerometer axes over more than 20 experimental acquisition sessions, containing about 10 fall events each, an algorithm has been developed. An example of the waveform output from the accelerometer is reported in Fig. 8. It is possible to note that a fall event is characterized by an acceleration peak, followed by some relatively constant values [11], and again followed in the experimental sessions by a peak due to the rise of the actor. The output data stream of each of the axes is analyzed, and an alarm is generated whenever at least one of the three axes detects a fall.

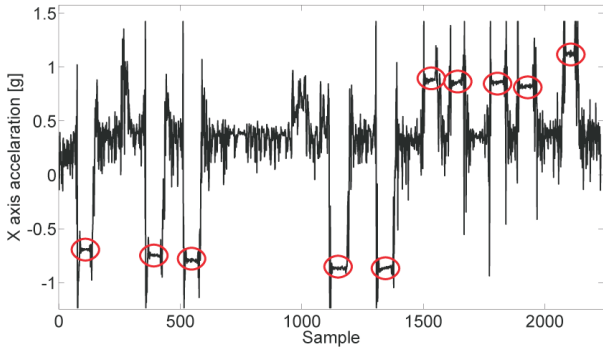


Fig. 8. X-axis acceleration data

The output bit-stream of each of the axes of the accelerometer is computed separately. A digital block compares the absolute value of the distances between every consecutive four samples to a threshold. When the sum is higher than the threshold, it means that the person has fallen. If the threshold is set low, some events, such as sitting down, can be confused with a fall event, while when it is set high some fall events are not detected at all. Since this algorithm is supposed to work together with additional sensors, and being practically impossible to detect all fall events with zero error rate, the above reported strategy has been improved. Three different values of the threshold have been used: TH_Low, TH_Medium, and TH_High. This method leads to three different alarm signals, each one described by the precision and recall reported in TABLE III that will

be considered in a different way by the request-for-help system gateway. The alarm generated by the circuit with TH_Low detects almost all possible events but generates also a considerable number of false alarms. By contrast, using TH_High, false alarms are very rare, but some fall events remain undetected. The TH_Medium specification constitutes a reasonable trade off. In the future, post processing of the accelerometer data together with the other sensors will permit to detect the maximum number of fall events, while minimizing the number of false alarms.

Table 3. ACCELEROMETER ALARM EFFICIENCY AND RELIABILITY

	Precision	**Recall**
Low threshold (TH1_L)	98.0%	56.0%
Medium threshold (TH1_M)	88.4%	79.3%
High threshold (TH1_H)	50.1%	96.2%

3.3 Acoustic signal analysis

For acoustic scene analysis, we have developed a statistical approach based on hidden Markov Models (HMMs). These HMMs are analogous to whole-word speech models, with a separate model used for each acoustic event class of interest. The HMMs [12] have a left-to-right topology consisting of 50 states to impose certain length constraints to the acoustic events. HMM training commences with a flat start that employs the Viterbi algorithm, using 13 dimensional Perceptual Linear Prediction (PLP) features as the front-end. Following training, recognition of the acoustic events employs an HMM network, similar in fashion to speech recognition decoding. An acoustic event is considered to be correctly detected, when its temporal center lies within the reference timestamps or vice versa. TABLE IV depicts the system performance under different decoding parameters, where the acoustic weight specifies how much the acoustic scores contribute to the final score. It turns out that he fall acoustic class gets

Table 4. AUDIO ALARM EFFICIENCY AND RELIABILITY

Acoustic weight	**Precision**	**Recall**
0.005	59.6%	59.1%
0.025	80.9%	42.2%
0.25	83.0%	35.2%

mostly confused with background and door slamming noises. There are lots of false alarms that contribute to the low recall depicted in TABLE IV. We plan to further improve performance, by using linear discriminant analysis (LDA) to differentiate falls from other sounds. In addition, we plan to experiment with

the maximum-a-posteriori (MAP) approach, which may be more suitable than maximum likelihood (ML) for HMM training for the particular problem at hand.

4 Conclusions

A hardware and software multi-sensory framework for high-recall fall detection has been presented. Three sensor data streams have been separately processed with suitable algorithms to maximize correct fall-detection performance. In future work, data fusion over the different sensors types will be studied, possibly based on a fuzzy logic decision model receiving as input variables the different data streams (audio, video and acceleration measurements). This feature is currently under development in order to improve overall precision and recall.

Acknowledgment

The presented framework has been developed within Netcarity consortium, funded by European Community. A special acknowledgment goes to MR&D Institute staff to A. Redaelli and A. Giacosi (Milan, Italy) for support and participation in data acquisition campaign and to Massimo Ferri, University of Pavia, for wireless accelerometer preliminary study.

References

1. N. Noury, A. Fleury, P. Rumeau, A. K. Bourke, G. O. Laighin, V. Rialle, J. E. Lundy: Fall detection Principles and methods. In: Proceedings of the 29th International Conference of the EMBS, IEEE (2007), 1663–1666
2. J. Chen, K. Kwong, D. Chang, J. Luk, R. Bajcsy: Wearable sensors for reliable fall detection. In: Proceedings of the 27th Engineering in Medicine and Biology Conference, IEEE (2005) 3551–3554
3. B. Jansen, R. Deklerck: Home monitoring of elderly people with 3d camera technology. In: Proceedings of the first BENELUX biomedical engineering symposium, Brussels, Belgium (2006)
4. http://www.netcarity.org
5. N. Ravi, N. Dandekar, P. Mysore, M. L. Littman: Activity recognition from accelerometer data. In: Proceedings of the 17th Innovative Applications of Artificial Intelligence Conference AAAI, IEEE (2005) 1541–1546
6. http://www.mesa-imaging.ch
7. http://www.freescale.com/files/sensors/doc/fact_sheet/MMA7260FS.pdf
8. http://www.shure.com/ProAudio/Products/WiredMicrophones/us_pro_MX391-C_content
9. D. Lee: Effective gaussian mixture learning for video background subtraction. In: Transactions on Pattern Analysis and Machine Intelligence 27 (5) (2005), 827–832
10. M. Isard, A. Blake: CONDENSATION - Conditional density propagation for visual tracking. In: International Journal of Computer Vision, 29 (1) (1998), 5–28
11. G. Brown: An accelerometer based fall detector: development, experimentation and analysis. Superb internship report, University of California, Berkeley (2005)
12. L. Rabiner: A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE, Volume 77 (1989) 257–286