An Active Vision System for Fall Detection and Posture Recognition in Elderly Healthcare

G. Diraco, A. Leone, P. Siciliano
CNR-IMM, Via Monteroni, presso Campus Universitario, Palazzina A3 – 73100 Lecce, Italy
[giovanni.diraco, alessandro.leone, pietro.siciliano]@le.imm.cnr.it

Abstract—The paper presents an active vision system for the automatic detection of falls and the recognition of several postures for elderly homecare applications. A wall-mounted Time-Of-Flight camera provides accurate measurements of the acquired scene in all illumination conditions, allowing the reliable detection of critical events. Preliminarily, an off-line calibration procedure estimates the external camera parameters automatically without landmarks, calibration patterns or user intervention. The calibration procedure searches for different planes in the scene selecting the one that accomplishes the floor plane constraints. Subsequently, the moving regions are detected in real-time by applying a Bayesian segmentation to the whole 3D points cloud. The distance of the 3D human centroid from the floor plane is evaluated by using the previously defined calibration parameters and the corresponding trend is used as feature in a thresholding-based clustering for fall detection. The fall detection shows high performances in terms of efficiency and reliability on a large real dataset in which almost one half of events are falls acquired in different conditions. The posture recognition is carried out by using both the 3D human centroid distance from the floor plane and the orientation of the body spine estimated by applying a topological approach to the range images. Experimental results on synthetic data validate the correctness of the proposed posture recognition approach.

Keywords—Fall detection, posture recognition, range imaging, self-calibration, plane detection.

I. INTRODUCTION

Fall detection for elderly has become an active research topic since the healthcare industry requires products and technology in this field. Many technical solutions have been studied and developed in the prevention/detection of falls and in the past two years a fall detection taxonomy has been defined according to the used sensors [1]. Fall detection methods are normally divided into three classes: wearable device-based, ambience device-based and vision-based methods. Cameras in in-home assistive system present several advantages over different sensors: they are less intrusive because installed on building (not worn by users), they are able to detect multiple events simultaneously and the recorded video can be used for post verification and analysis. According to the used principles relating to the characteristics of fall, vision-based approaches can be divided into three categories: inactivity detection, shape change analysis and 3D head motion analysis. A detailed review about the vision-based approaches can be found in [2]. Since spatial information is of direct relevance for both posture classification and fall detection and since it is not possible to obtain reliable spatial information by using mono-camera vision systems (or omni-camera vision systems for ceiling-mounting setup [3]), multi cameras approaches (including the stereo vision techniques) have been used [4] but they appear computationally expensive providing poor results in uniform regions or in severe illumination conditions. A practical problem is the calibration activity that relates the camera coordinate system to a more useful coordinate system. Normally manual or semi-automatic calibration approaches [5] are used.

This paper presents a framework for the monitoring of ageing people, including fall detection capabilities, by using a self-calibrated vision-based system without any user intervention. To overcome the previously discussed drawbacks of the passive vision, a Time-Of-Flight (TOF) 3D camera is employed. It is important to note that the proposed active vision system allows to guaranty the privacy of the people since chromatic information is not available (only target depth is processed). On the other hands, the spatial resolution of the active sensor is inappropriate for details acquisition such as nose, mouth and eyes in the face, so that the identity of the person is preserved. Moreover, if the privacy issue is more strict the fall detection device can be configured to trigger in output only a fall/non-fall signal. Moreover, a preliminary study on posture recognition is proposed according to a 3D hybrid approach in which model-based and learning-based techniques are used in conjunction in order to estimate location and orientation of the different body articulations (more details about the posture recognition taxonomy can be found in [15]). A Discrete Reeb Graph-based body skeleton [18] is extracted and the Geodesic distance is used as Morse function [20] since it provides a better segmentation of the human body, allowing the invariance to rigid transformation (translation, scale and rotation) and isometric deformation [17].

Section 2 presents the setup of the active vision system, underlying the main characteristics of the used 3D camera. An automatic floor detection strategy is presented in Section 3, accomplishing the off-line self-calibration procedure, whereas Section 4 describes people segmentation process by applying...
well-known techniques on raw depth information. Once segmentation is completed, 3D position of the elderly in the room (or more properly its centroid) is analyzed and a fall event is detected by thresholding the distance of the centroid from the floor (Section 5). Experimental results show fall detection performances in terms of reliability and efficiency on a large real dataset acquired in quasi-real conditions. A qualitative analysis for the posture recognition approach is shown in Section 6 underlying the key components of the proposed methodology.

II. ACTIVE VISION SYSTEM OVERVIEW

In the last years several active vision systems 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, range cameras present several advantages in the use (i.e. small dimensions, low power consumption, etc.), integrating distance measurement as well imaging aspects (RIM - Range Imaging). An overview of the active vision system is described in this section. The stationary TOF camera is installed in a wall-mounting configuration (Fig. 1), assuming that:

a) \( H \) is the distance of the camera optic center from the floor plane;

b) \( \theta \) is the tilt angle (i.e. the angle between the floor plane and the z-axis);

c) \( \beta \) is the roll angle that can be hold as negligible in this setup (i.e. the rotation around the z-axis is almost null).

Due to the limited Field-Of-View (47.5×39.6 degs) of the camera, the static wall-mounting configuration has been preferred to the ceiling-mounting one. The main characteristics of the used active sensor are discussed in the following subsection. Moreover, advantages in the TOF camera usage are presented and the proposed solution is compared with a generic camera, the static wall-mounting configuration has been installed in a wall-mounting configuration (Fig. 1), assuming that (i.e. the rotation around the z-axis is almost null).

Figure 2. The \( \alpha_i \) and \( \alpha_j \) angles between the normal vector \( \vec{n}_i \) of the plane \( \pi_i \) and the coordinate axes are used for planes filtering.

III. SELF-CALIBRATION BY AUTOMATIC FLOOR DETECTION

Since the 3D points from the TOF camera are represented in a coordinate system centered in the camera optic center, a coordinate changing is needed to represent metric measures in the world coordinate system, allowing to describe in a simple way the 3D position of the elderly in the world (i.e. the room). The coordinate system changing requires the knowledge of the transformation matrix from the camera to the world coordinate system, including orientation and position of the camera (i.e. external camera calibration). In monitoring applications (including fall detection systems), the camera positioning and orientation may differ from installation to installation, so that a self-calibration is presented. The proposed calibration method retrieves the camera orientation and position automatically, without using reference frame, calibration tool [9] (i.e. calibration pattern, landmarks, etc.) or user intervention [5]. Since a fall event is detected by evaluating the distance of the centroid of the people from the floor in the pan plane (the yz-plane), the coordinate transformation can be thought as a rotation around the x-axis (\( \theta \) angle), a rotation around the z-axis (\( \beta \) angle) and a translation (\( H \)) with respect the floor plane (see Fig. 1). The relation between camera and world coordinate systems is described by the following equation:

\[
\begin{bmatrix}
X \\
Y \\
Z \\
\end{bmatrix} = \begin{bmatrix}
\cos \beta \sin \theta & \cos \theta & \sin \theta & 0 \\
\cos \beta & -\sin \beta & 0 & 0 \\
\sin \beta \sin \theta & \cos \beta \sin \theta & -\cos \theta & H \\
0 & 0 & 0 & 1 \\
\end{bmatrix}\begin{bmatrix}
x \\
y \\
z \\
1 \\
\end{bmatrix}
\]

(1)
where \((X,Y,Z,1)\) and \((x,y,z,1)\) are the homogeneous coordinates in the world system and in the camera system, respectively. The detection of the external calibration parameters \((0, \beta, H)\) is accomplished during the first installation of the device by searching the calibration plane (i.e. the floor) or whenever the device is moved. Assuming that the camera is always looking to the floor, the calibration plane is detected by a two-steps strategy. The first step deals with the detection of enough large planes in the 3D scene, whereas in the second step the parameters that characterize the detected planes are filtered out by providing the floor plane and the external calibration parameters. In particular, the planes detection procedure searches iteratively the biggest plane in the 3D points cloud and removes those that belong to the detected plane. For a given iteration, the algorithm works with a subset of the 3D points used in the previous iteration. The procedure finishes when no more points can be fitted on a plane or when a prefixed amount of iterations is achieved. Since 3D points of a scene are normally affected by noise due to scattering problems in the acquisition, planes are detected by using a RANSAC-based approach [10] which is robust in the sense of good tolerance to outliers in experimental data. Let the \(i\)-th iteration of the algorithm, the RANSAC plane detector provides the four parameters \((a,b,c,d)\) that describe the implicit model of the fitted plane \(\pi^i\): 
\[
a_ip_x+b_ip_y+c_ip_z+d_i=0
\]
with \(p=(p_x,p_y,p_z)\) a generic point belonging to the detected plane \(\pi^i\) in camera coordinates. For each detected plane, trigonometric considerations allow to derive the \(\alpha^i, \alpha'^i\) and \(\alpha''^i\) angles between the normal vector \(\vec{n}_i=[a_i, b_i, c_i]^T\) of the considered plane \(\pi^i\) and the coordinate axes (Fig. 2), that is:
\[
\alpha^i = \arccos \frac{a_i}{||\vec{n}_i||}, \quad \alpha'^i = \arccos \frac{b_i}{||\vec{n}_i||}, \quad \alpha''^i = \arccos \frac{c_i}{||\vec{n}_i||}
\]
(3)
Since the measurements are accomplished in the pan plane, constraints on \(\alpha''^i\) angle are not required. The \(\alpha'^i\) and \(\alpha''^i\) angles are now filtered according to appropriate constraints to determine planes that are compatible with a floor plane. Since for an optimal installation the \(\alpha''^i\) angle should be closed to 90\(^\circ\) (i.e. the \(\beta\) angle is almost null), a \(\pm10^\circ\) of tolerance is accepted so that the constraint for \(\alpha'^i\) is \(80^\circ \leq \alpha'^i \leq 100^\circ\). Similarly, a reasonable constraint for the \(\alpha''^i\) angle is \(50^\circ \leq \alpha''^i \leq 80^\circ\), although the appropriate choice deals with the specific distance \(H\) between the camera and the floor plane. For each plane \(\pi_i\) filtered by the previous activity (i.e. floor plane, table surface, chair surface, etc.), the centroid \(C_i\) of the 3D points belonging to \(\pi_i\) is computed and the corresponding projection onto the pan plane (i.e. \(C'_i\) in Fig. 2) is used for the floor plane selection. In particular, among the \(m\) filtered planes, the algorithm selects as floor the plane \(\pi_m\) that satisfies the relation:
\[
F = \arg \max_{i \in \text{floor}} d_i \quad \text{with} \quad d_i = ||C'_i - O||
\]
(4)
Finally, the calibration parameters \((0, \beta, H)\) are determined as \(0=\alpha'^i, \beta=\alpha''^i , 90^\circ, H=d_p\). The distance \(h\) of the people centroid from the floor plane is the feature used in the fall detection. By applying the eq. (1), it can be expressed as:
\[
h = \sin \beta \sin \theta \cdot c_x + \cos \beta \sin \theta \cdot c_y - \cos \theta \cdot c_z + H
\]
(5)
where \(\{c_x,c_y,c_z\}\) is the people centroid in camera coordinates. The above eq. (5) is valid for whatever 3D point of the scene. The self-calibration algorithm has been evaluated in a home-like environment that presents the floor plane and many planar surfaces. A MEMS-based Inertial Measurement Unit (IMU) [11] attached to the TOF camera has been used to derive ground truth data. The IMU sensor provides drift-free 3D orientation with a static accuracy lower than 0.5°, allowing to measure the external calibration parameters carefully \((\theta=78.2^\circ, \beta=3.4^\circ, H=2.05m)\) are referred to the scene shown in Fig. 3, in which the 32% of the whole amount of 3D points of the scene are floor points). Calibration results are shown in Fig. 3: the first algorithmic step detects a wall plane (Fig. 3.c), a table plane with some chairs planes (Fig. 3.e) and the floor plane (Fig. 3.d). In the second algorithmic step the floor plane is correctly selected (the green plane in Fig. 3.d), providing an accurate estimation of the calibration parameters \((0=78.2^\circ, \beta=3.4^\circ, H=2.05m)\).
the instability of that pixel. In order to extract quantitative information about the person in the scene and determine the distance of its center-of-mass (the 3D centroid) from the floor plane, a Bayesian segmentation is used. The raw depth map (i.e. the metric z-coordinate provided by the sensor) is used for the background modeling. Since the depth measure is affected by fluctuations due to high object reflectivity, ambiguity range presence, etc., a robust modeling method is required. The Mixture of Gaussians (MoGs) method proposed in [12] has been implemented. Compared with the traditional formulation [13], this algorithm improves the convergence rate without compromising model stability, by replacing the global and static retention factor with an adaptive learning rate calculated for each Gaussian at every frame. The approach ensures that the same effective learning for each Gaussian component is applied throughout all stages of the system, including reassignment and system reset. In the implementation, the background model is trained with K=3 Gaussians whereas 0.005 is the learning parameter of the adaptation. The background model is updated online and it is used for the detection of moving objects through a Bayesian segmentation [12]. Fig. 4 shows segmentation on depth information (Fig. 4.c, Fig. 4.d). The segmented binary image is filtered by morphological operators (erode and dilate with 3×3 structuring element) and the connected components are used to refine the segmented blobs. By thresholding the aspect ratio of each connected component, blobs not related to human silhouettes can be discarded.

Once people silhouette has been detected in the 3D world and its center-of-mass has been computed, a tracking strategy allows to link people silhouettes in different time instants. A widely used approach for tracking is the Kalman filter applied to each segmented object. This approach requires the use of a high complexity management system to deal with the multiple hypotheses necessary to track objects. Due to the non-linear nature of human motion, in this context we adopt a stochastic approach which is based on the ConDensation algorithm (Conditional Density Propagation over time [14]) that is able to perform tracking with multiple hypotheses in range images (500 samples are used for people tracking). 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. The 3D centroid is predicted frame-by-frame in the range data, according to a state vector defined by merging the following information: \((x;y;z)\)-coordinates of the centroid and the corresponding speeds along the \((x;y;z)\)-axes. The tracker is realized by thresholding the Euclidean distance between the predicted location of the centroid and its measured version in the adjacent time step. Heavy situations as the partial occlusions are correctly detected by evaluating the distance of the lower part of the segmented people from the floor plane according to the eq. (5), allowing to adjust the position estimated by the particle filter. Fig. 4.e shows the result of the tracking in real home conditions.

V. THRESHOLDING-BASED FALL DETECTION AND RESULTS

A fall event is detected by thresholding the distance \(h\) of the 3D centroid from the floor plane (see eq. (5)). In general a fall event is characterized by:

1) a distance of the centroid from the floor plane (the feature \(h\)) lower than a prefixed value;
2) an unchangeable situation (negligible movements of the people) for at least 4 seconds.

Fig. 5 shows the \(h\) trend when fall events occur (red circles). The fall detector application has been implemented in C++ on a Intel Core Duo Processor, providing real time performances (up to 25fps). A set of scenarios for the evaluation of the proposed fall-detector has been used. More than 450 events have been acquired and approximately 210 falls have been simulated in different conditions (backward falls, forward falls, lateral falls, etc.). The performances have been defined according to the criterion of quality described in [1] (efficiency and reliability). Table I shows the performances of the fall detection algorithm for different values of the threshold (0.4 meters provide the best choice to limit false

![Figure 4](image4.png)

**Figure 4.** The segmentation (d) is accurate on raw depth map (c) (depth is in the [0, 255] range for visual purpose). The black cross (e) is referred to the predicted position of the 3D centroid according to the tracking algorithm. In white the measured 3D centroid (cross) and the bounding box of the blob (rectangle). The segmentation process is corrupted (b) when intensity information is used (a).

![Figure 5](image5.png)

**Figure 5.** Falls (red circles) are characterized by a distance of the people centroid from the floor plane lower than a threshold.
alarms). Total occlusion situations (i.e. the person falls behind a big object) and poor segmentation during the transient period of the background modeling affect the performances.

### TABLE I. FALL-DETECTOR EFFICIENCY AND RELIABILITY

<table>
<thead>
<tr>
<th>Threshold for centroid height</th>
<th>Efficiency (%)</th>
<th>Reliability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 meters</td>
<td>51.1</td>
<td>99.2</td>
</tr>
<tr>
<td>0.4 meters</td>
<td>80.0</td>
<td>97.3</td>
</tr>
<tr>
<td>0.5 meters</td>
<td>81.3</td>
<td>89.3</td>
</tr>
</tbody>
</table>

**VI. 3D GEODESIC DISTANCE MAP-BASED POSTURE RECOGNITION: A PRELIMINARY STUDY**

The posture recognition is carried out by using two kind of features: the distance of the human centroid from the ground floor and the orientation of the human spine obtained by an appropriate extracted 3D skeleton. The Discrete Reeb Graph (DRG) [18] is used for the skeleton extraction. The skeletonization by Medial Axis Transform (MAT) is usually a good choice in 2D, but it can be proven that MAT of a 3D object consists of 2D smooth surface patches and 1D axes [19]. Moreover, the Reeb Graph is a skeletal graph that encodes a 3D surface by using a 1D representation and hence it is more suitable for analytical posture representation. Each node of a Reeb Graph represents a level-set curve on a 3D surface defined by a real function \( f \). The function \( f \) is a Morse function [19] when the critical points on the 3D surface are not degenerate. Morse theory can be considered as a generalization of the classical theory of critical points (maxima, minima and saddle points) of smooth functions. Morse theory states that for a generic function defined on a closed compact manifold (e.g. a closed surface) the nature of its critical points determines the topology of the manifold. Reeb Graph shows graphically the configuration of critical points and their topological relationships. Hence, the Reeb Graph gives a compact description of the evolution of level set curves underlying the topological structure. In recognition applications the main aspect related to Reeb Graph extraction concerns the invariance under some transformations that can be achieved by defining a proper Morse function. Among all invariant Morse functions proposed in literature [20], in this work the Geodesic distance has been used since it is invariant to translation, scale, rotation and isometric transformations and it ensures high accuracy in body parts segmentation [17].

The Reeb Graph is extracted by applying the DRG which is an extension of the classical method to unorganized cloud of 3D points. The DRG method is suited to deal with dense 3D representation (i.e. connected mesh) of closed surfaces. Since range image provides only 2.5D information, the DRG must be adapted to handle range data [16]. In this work, the DRG adaption on range data is achieved by detecting and interpolating overlapped regions in order to generate a connected mesh. In particular, the bounds of overlapped regions (arm and leg in Fig. 6.a) are detected by applying the Laplace operator to the depth map (Fig. 6.b). Once the overlapped region bounds are detected, the level-sets of the height function (depth map) allow to label the inbound regions (Fig. 6.c) and to interpolate them (Fig. 6.d). After the interpolation step, the connected mesh can be defined and the DRG is calculated from the geodesic distance map by using the method described in [18]. The four main postures (lie, sit, stand, bend) used in the study are shown in Fig. 7.a. The related geodesic maps and Reeb Graph-based skeletons are shown in Fig. 7.b and Fig. 7.c, respectively (the colour in Fig. 7.b describes the normalized distance of each point of the 3D mesh from the body centroid - blue is related to small distance,

![Graph 6. Example of detection and interpolation of overlapped regions](image)

![Graph 7. The four main postures](image)
been validated by using synthetic range data (generated with the commercial modelling software Poser) in order to have a ground-truth. Posture discrimination is evaluated by plotting the centroid distance on the y-axis. The 4 clusters are qualitatively well separated, showing the goodness of the selected features. The presence of 4 well defined clusters proofs the validity of the proposed approach for the recognition of postures.

VII. CONCLUSIONS

This paper presents a method for fall detection in 3D range image that combines information about the 3D position of the centroid of the people with the detection of inactivity. This information is computed through a self-calibration algorithm able to detect external calibration parameters automatically. The procedure does not require landmarks, calibration patterns or user intervention so its application is useful in monitoring and surveillance applications. Well-known techniques for background modeling, people segmentation and tracking have been used and the obtained performances demonstrate the goodness of the proposed method in a real-time application. A novel approach for posture recognition is proposed exploiting the usefulness of topological information for posture discrimination. A preliminary study based on 4 main postures and synthetic range data allows to validate the approach. Future works are addressed to recognize more complicated behaviors of elderly, even in conjunction with other kind of sensors (i.e. microphone, accelerometer) in a multi-sensorial approach in order to improve the reliability of the system. The future activities dealing with posture classification are oriented to validate the proposed approach by using real range data and to investigate more suitable classification strategies.

REFERENCES


