Human activity analysis: a personal robot integrating a framework for robust person detection and tracking and physical based motion analysis.

Abstract

The analysis of certain parameters related to cognitive and to motor humans’ activities in everyday life conditions can allow to detect potential behavioral troubles, make diagnoses and assess patients’ progress after a therapy. Within this context, personal robots can provide an autonomous movable platform for embedded sensors allowing to detect and track humans while ensuring an optimal observability of the person’s activity in complex and cluttered environments. This paper presents a framework combining a multimodal human detector based on sensors embedded in a mobile robot and a decisional engine exploiting the fuzzy logic mechanisms to make the robot track humans, maximizing observability and facing losses of detection. The robustness of this framework is evaluated experimentally in home spaces through different scenarios. Such a mobile system provides an effective marker-less motion capture means for sensing human activity in non-invasive fashion. We present a physical model based method exploiting the features of the system and of the embedded Kinect. Its performances are evaluated first comparing the results to those obtained with a precise 3D motion capture marker based system and to data obtained from a dynamic posturography platform. Then an experiment in real life conditions is performed to assess the system sensitivity to some gait disturbances.

Keywords

Human activity analysis · human detection · human following · marker-less motion capture · dynamic human body modeling

© 2013 Consuelo Granata et al., licensee Versita Sp. z o. o.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs license, which means that the text may be used for non-commercial purposes, provided credit is given to the author.

1. Introduction

The analysis of human activity in everyday life contexts using an autonomous robotic system can afford to consider non-invasive implementation of means for the quantification of some executive functions. Such a system can be used, for example, to monitor behavioral tasks and goal-oriented exercises or to detect the loss of spontaneity and responsiveness to external events, that are both indices of neurodegenerative diseases (e.g. Parkinson and Huntington’s diseases). Stimulating the activity of individuals or observing them achieving tasks in some slices of daily life suggests to design and develop mobile autonomous robots able to search and detect humans, to appropriately interact with them and to collect data useful for the analysis of their behavior. The main abilities that this kind of robot has to fulfill are to detect, localize and track humans and analyze their activities. These capabilities are particularly challenging in real life situations, especially if only sensors embedded on the robot are employed. The real added value of such a system, is the possibility to maximize the observability of the person’s activity without limiting the person’s movements. The existing systems for activity analysis exploit static sensors (e.g. IR and RGB cameras) or instrumented surfaces (i.e. treadmills and force platform) entailing observability problems or some inevitable restrictions of the workspace. On the contrary, the use of a mobile platform can maximize the observability, provided the robot is able to detect, track and follow a moving person. This approach involves challenging difficulties arising from occlusions, obstacles and navigation problems. Reliable person following behavior in indoor cluttered environments is particularly challenging. Solving this problem in full autonomy requires the robot to perceive and localize fixed and moving objects in the environment (context), estimate the state of the person, predict his movements and anticipate his path if necessary. Following processes based for example merely on the control of the distance [1] and/or of the relative speed [2] with respect to the person have shown limitations in real life environments. A reliable following behavior can be produced only by exploiting appropriate reasoning mechanisms able to take into account a large number of data (concerning the person’s state, the environment and the task to achieve) and to deal with uncertainties (the information available via the robot perception is often scarce, vague, inconsistent and incomplete). The advantage of making use of a mobile robot to track human activities by embedded sensors is in particular the possibility to control the conditions of information retrieval, to focus resources, to reduce the risk of occlusions and the measurement noise. This allows an in-depth analysis in non controlled environmental conditions, without any environmental equipment required. Accessing spatio-temporal parameters of human activity, analyzing and interpreting them at different time and space scales are very interesting issues that attract a large num-

*E-mail: granata@isir.upmc.fr
†E-mail: philippe.bidaud@isir.upmc.fr
‡E-mail: salini@isir.upmc.fr
§E-mail: ady@isir.upmc.fr
number of contributions. This great interest is pushed by a broad range of applications in many areas such as virtual reality, advanced user interfaces, disease diagnosis and more generally, the understanding of human behaviors. The analysis of human activity can be grounded on different motion capture technologies such as wearable sensors, instrumented platforms, optical markers and cameras. In the past few years, several studies focused on reliable human movements tracking by employing cameras. Commonly, the video stream is exploited through filters such as particle filter [3, 4]. Hybrid Monte Carlo filter [5] and Smart particle filter [6] to retrieve and track some kinematic characteristics and global/local optimization [7–9] is used to compute the best global/local estimation of the body pose. More recent methods are model based [10]. These use a human body model, which includes kinematic structure, body appearance and motion. A function that matches the local measurements (image domain) with the model (reproducing through a skeleton the geometric constraints imposed upon the movement), is used. Rarely, physical models are used as well [11]. The advantage of introducing a physical model is to have richer information about internal and external efforts and more generally to track the reaction of the human motion for accurate analysis of behaviors and motor activities. For more details about the motion capture methods we refer to the Poppe’s review [12].

The paper is organized as follows. In Section 3, the experimental robotic system used for developments and experiments is described. Section 3 introduces the framework for the synthesis of smart person following behavior composed of a multimodal human detector and a fuzzy logic based decisional engine. The robustness of the framework is experimentally evaluated in cluttered environments based on different scenarios. The system developed for physical based motion analysis is detailed in Section 4. A method for determining the individual’s anatomical parameters and the algorithms developed for the digital animation model based on the measurement of a number of characteristic points obtained by a Kinect embedded in the robot are described. These algorithms are validated experimentally on different exercises during which the positions of the Center of Pressure (CoP) measured and calculated using a posturography platform, are compared. The overall proposed system is assessed analyzing the gait of a subject followed by the robot. The system shows strong sensitivity to some disturbances obtained manipulating artificially the lower limb joints. In Section 5 conclusions are provided.

2. Experimental system

The robot used for the implementation is a commercial robot, Kompaï (Figure 1) specifically designed for services in private or public spaces. It is based on a differential nonholonomic mobile platform equipped with a Sick-5300 laser and a webcam (embedded in its head at about 1.2 m above the ground). The low-level control system is managed by the embedded Microcontroller (Emtrion SH7780 SBC). The high-level control system is integrated in a tablet PC running Windows 7 OS and is programmed in MRDS (Microsoft Robotics DeviceStudio). A SLAM (Simultaneous Localization and Mapping) software is exploited for the robot motions. In order to raise the potential of this robot, we added a Microsoft Kinect and an extra laptop for computation. We have integrated the SpirOps software for the decisional engine (see Sec. 3.1.2 for more details).

3. Framework for person following behavior

The implementation of service robots should be supported by a decisional layer capable of organizing basic activities, depending on their functional components and on their specific tasks, but also in relation to the requests of the users and to the interaction with them. A social robot needs reactive behaviors which adapt to changing context conditions in presence of uncertainties. The robot behavior must be appropriate to the current context (environment, type of service) and the user profile (needs, activity). This corresponds to reproducing a real reasoning process that has to deal with information scarce, vague, inconsistent or incomplete because of robot perception limitations and of the hard environmental conditions. We propose a framework for robust person following based on 1. A multimodal human detector which exploits different low-cost sensors embedded in the robot (Section 3.1.1); 2. A decisional engine which processes the robot perception to produce a reliable person tracking (Section 3.1.2). This framework has been implemented on the Kompaï robot and its efficacy is demonstrated in different social relevant scenarios, including long-range as well as face-to-face situations (Section 3.2). Through these scenarios we also show how the robot is able to achieve interaction and motion tasks in cluttered environments, even in case of noisy and uncertain perception of the person and of the surroundings.

3.1. Methods

3.1.1. Human detector

Human detection plays a crucial role in many application areas, especially for mobile social robot working in scenarios involving human motion. Different features specific to humans can be considered for the detection such as full or partial body shape, face texture, voice and movements among others. Accordingly, different kind of sensors can be employed. Several research papers focused on vision systems that only operates a visible spectrum camera (e.g. 2D color image) [13, 14]. Integration of different perception channels (multi-modal perception) seems mandatory to obtain an overall richer information for more reliable results. For human detection, the most common solution combines a laser sensor and a camera used respectively for legs detection and for face or body detection. The appropriate combination of laser scanners and cameras is particularly suitable because the first ones provide accurate but poor information, conversely the second ones offer discriminative but imprecise data. Many reports [15–18] aimed to coupling laser legs detection and camera face detection. In [19] the authors introduced a probabilistic approach for tracking people which realizes an equitable fusion of different sensory systems: a laser and a sonar system for legs detection, and a fisheye-based omni-directional camera for skin color detection. All these works suggest interesting methods which however are effective only if the person’s face is visible to the camera. In [20] Zhang used a laser to define regions of interest by focusing moving objects. This method has the advantage of reducing the computational time but it is not very robust in the case of static persons. Other multimodal perception systems for human detection are based on the combination of visual and auditory sensors [21–23]. In recent years, Photonic Mixer Devices (PMD) camera has received increasing attention because it is a powerful device able to provide three different maps trough different
modalities (i.e., depth map, modulation amplitude map and intensity image) by exploiting the Time-of-Flight (ToF) principle. With the arrival of the Microsoft Kinect, the combination of standard color camera and depth sensor that captures video data in 3D has spread greatly. With the software launched in 2012 by Microsoft, new features are provided by the Kinect: skeletal tracking enhancement, advanced speech and audio capabilities and improved synchronization between color and depth among others. Our human detection system grounds essentially on the complementarity of a laser, a RGB camera and a Kinect. Four detectors are employed: legs detector, face detector, upper body detector and sound source detector.

1. Laser based legs detector

For the legs detection we propose a method aimed at recognizing the geometrical signature of legs pair. The laser sensor is used to detect the legs of the persons assuming that they are standing or seated. First we have modeled the legs size and posture (distance between 2 legs) by the means and the standard deviations (in m) given in \( \mu, \sigma \).

\[
(\mu, \sigma)_{\text{radius}} = (0.13, 0.03) \\
(\mu, \sigma)_{\text{posture}} = (0.23, 0.04)
\]

Using these models, the legs detector follows 3 steps. 1. We first do a scan segmentation: we look at points of interest by detecting “blobs” in the laser scan. A blob is a set of consecutive laser reflection points supposed to belong to the same object. Two points belong to the same blob if their distances are in the range of 0.1 m. 2. For the recognition of a blob as leg, we apply the defined Gaussian model \((\mu, \sigma)_{\text{radius}}\) (the used recognition threshold is \(-3\sigma_{\text{radius}}\)). 3. Two legs are considered as a pair of legs if their distance is in the Gaussian model \((\mu, \sigma)_{\text{posture}}\) (the used recognition threshold is \(-3\sigma_{\text{posture}}\)). The detector shows experimentally a correct detection rate > 30% within 3 m (with a laser scan frequency of 12.5 Hz, this means an average of 4 detections per second). From a distance of 3 m, the legs detector performances begin to drop: with 0.5° scan resolution the probability for laser rays to intersect a leg decreases drastically with distance.

2. RGB camera + laser based face detector

Many different algorithms exist to perform camera based face detection (e.g. using skin colors, contours or other complex features), but the recurrent problem in applying them is the required computational time \([24]\). Viola and Jones \([25]\) developed a method to detect faces within an image in an accurate and fast fashion. Their algorithm is based on the combination of classifiers in a cascade, which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. Through this regionalization the detection speed is considerably increased due to the reduction of the size of the investigated area. Our face detector is based on this method: we first use the ADAboost classifier cascade library provided from OpenCV 2.1 to identify rectangular regions containing faces in the image. The reliability of the detector is then enhanced by applying the
particle filter proposed by Brocklehurst [26]. At each iteration we first draw a set of 100 points according to a Gaussian distribution centered on the detected face. We then compute an RGB histogram for each sample. The correlation between the histogram built for each sample and the histogram built for the face when it was first detected is computed on the 3 dimensions (RGB) by using the Pearson’s correlation coefficient (Eq. 2).

$$r(H_1, H_2) = \frac{\sum (H_1(l) - \bar{H}_1)(H_2(l) - \bar{H}_2)}{\sqrt{\sum (H_1(l) - \bar{H}_1)^2 \sum (H_2(l) - \bar{H}_2)^2}}$$

(2)

where

$$\bar{H}_1 = \frac{1}{N} \sum H_1(l)$$

(3)

and N is a total number of histogram bins. The similarity between 2 histograms is computed with the weighted sum (4).

$$s = w_r * r_r + w_g * r_g + w_b * r_b$$

(4)

where $r_r$, $r_g$ and $r_b$ are the correlation coefficient for the 3 colors (red, green and blue) and the weights $w_r$, $w_g$ and $w_b$ are chosen to give more importance on red, which is the skin predominant color ($w_r = 0.4$, $w_g = 0.3$ and $w_b = 0.3$). Therefore, the final weight $w$ associated to each sample is computed by using

$$w = e^{-a(1-s)}$$

(5)

The parameter $a$ is used to adjust the rapidity of the final weight drop as similarity decreases (we consider $a = 15$). The sample with the highest weight is considered as new coordinates of the tracked face on the image. Due to the fact that with visible spectrum camera the depth information is missing, the 3D position of the detected face is still unknown. The distance of the face to the robot is obtaining by matching the angular position given by the embedded laser sensor. The proposed face detector shows good results for frontal (a rotation up to 60° is tolerated) and near the camera (closer than 3 m) faces. This allows one to detect the face and upper body detectors. In the environment there are several objects with a shape similar to legs (such as tables and chairs) one can decide to penalize the legs detector with a weight lower than the others detectors. If the environment is not well lighted it is more convenient to penalize the face and upper body detectors. In zones 3 and 5, $c_f(p)$ has to be always null because the legs detector doesn’t work there. The merging zones for each detector are listed in Table 1. As the detectors work in asynchronous way, we manage the fusion considering their timestamps. The resulting $c(p)$ is a value between 0 and 1.

The state of the person (position and velocity) cannot be determined by direct observation due to lack of detection, to false positives and noise in measurements. To determine the person movements and to predict its 2D trajectory, the information provided by the person detector is processed by an Extended Kalman Filter (EKF) which integrates a human locomotion oriented prediction by relying on Hitcheur’s study [35]. For more details about the used model we refer the reader to [36]. The use of the human state predictor has two advantages. First, predicted trajectories are smoother than the detected trajectories (aberrant detections and noises are filtered). Second, it solves the problem of lack of detection in a lot of complex situations.

3. Kinect based upper body detector

The RGB and depth cameras, integrated in the Kinect, are used for upper body detection and tracking. The same classifier employed in face detection is used (ADABosst classifier cascade from OpenCV 2.1), but is trained on the upper body. The distance is taken as the mean distance computed across the area of the whole upper body. The performance of our detector is slightly lower than that of face detection. This is explained by the regularity of the facial structure, in comparison to the upper body which can be distorted by clothing and variability in size and shape.

4. Kinect + laser based sound source detector

A sound source detector can help in detecting the presence of people producing sounds (e.g. clapping hands or speech). We use the array of microphones integrated with the Kinect combined with the HARK software developed at Kyoto University (Japan). HARK provides various signal processing modules ranging from sound source localization and separation [27]. The source localization module uses Multiple Signal Classification (MUSIC) to localize multiple sources of sound in a robust manner. The source separation module uses an adaptive frequency-domain blind source separation algorithm known as GHDSS (Geometrically constrained High-order Decorrelation based Source Separation). The system is able to locate sounds with an accuracy of 5° (the performance degrades when the sources are close together). The distance of the sound sources from the robot is obtained by matching the angular position given by HARK with measurements from the laser sensor.

5. Fusion

The information extracted from the four detectors is merged by using a weighting criteria in accordance with the Field of View (FoV) (Figure 2) and the respective performances of each sensor (see Figure 3 and Table 1). The employed weighted sum is given by

$$c(p) = c_f(p) \ast dLegs(p) + c_f(p) \ast dFace(p) +$$
$$+ c_f(p) \ast dBody(p) + c_f(p) \ast dSound(p)$$

(6)

where $c_f(p)$ is the confidence value associated to the detection of a person in the 2D point p (we project to the floor the results of the 4 detectors), $dLegs(p), dFace(p), dBody(p)$ and $dSound(p)$ are binary values indicating if the detectors (respectively legs, face, body and sound detector) have detected a person or not in p (with a tolerance of 0.1 m). $c_f(p)$, $c_l(p)$, $c_b(p)$ and $c_s(p)$ are the weights associated to the different detectors. These weights are defined according both to the type of environment and to the detection zone (Figure 3) where the point p is located. For instance, if in the environment there are several objects with a shape similar to legs (such as tables and chairs) one can decide to penalize the legs detector with a weight lower than the others detectors. If the environment is not well lighted it is more convenient to penalize the face and upper body detectors. In zones 3 and 5, $c_f(p)$ has to be always null because the legs detector doesn’t work there. The merging zones for each detector are listed in Table 1. As the detectors work in asynchronous way, we manage the fusion considering their timestamps. The resulting $c(p)$ is a value between 0 and 1.

The state of the person (position and velocity) cannot be determined by direct observation due to lack of detection, to false positives and noise in measurements. To determine the person movements and to predict its 2D trajectory, the information provided by the person detector is processed by an Extended Kalman Filter (EKF) which integrates a human locomotion oriented prediction by relying on Hitcheur’s study [35]. For more details about the used model we refer the reader to [36]. The use of the human state predictor has two advantages. First, predicted trajectories are smoother than the detected trajectories (aberrant detections and noises are filtered). Second, it solves the problem of lack of detection in a lot of complex situations.

3.1.2. Decisional engine

The human detection and trajectory tracking in cluttered and non-structured environments are challenging issues because of environmental noises, occlusions, sensors limitations, algorithmic mistakes, and so on. An improvement can be obtained by exploiting the mobility of the robot in order to place the embedded sensors in the optimal position for sensing the human, which corresponds to a real decision problem. The use of precise mathematical models in a situation where
Table 1. Sensors used for human detection: merging zones (Figure 3) where their result is taken into account, detectors based on them and FoVs in traverse and sagittal planes.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Merging zones</th>
<th>Detectors</th>
<th>Traverse plane FoV</th>
<th>Sagittal plane FoV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laser</td>
<td>1, 2, 3, 4</td>
<td>Legs*, Face, Sound</td>
<td>270°</td>
<td>0°</td>
</tr>
<tr>
<td>RGB camera (webcam)</td>
<td>1, 2</td>
<td>Face</td>
<td>78°</td>
<td>78°</td>
</tr>
<tr>
<td>RGB camera (Kinect)</td>
<td>2, 3</td>
<td>Upper body</td>
<td>57°</td>
<td>43°</td>
</tr>
<tr>
<td>Depth camera</td>
<td>2, 3</td>
<td>Upper body</td>
<td>57°</td>
<td>43°</td>
</tr>
<tr>
<td>Microphones</td>
<td>1, 2, 3, 4, 5</td>
<td>Sound</td>
<td>360°</td>
<td>360°</td>
</tr>
</tbody>
</table>

* Legs detector works only in zones 1 and 2.

informations are scarce, vague, inconsistent, or incomplete could lead in many cases to inappropriate decisions and behaviors. The problem of decision under uncertainties (missing data, imprecise data) and for vague or fuzzy objective or multi-objectives (also conflicting), can be addressed by fuzzy decision making. Fuzzy logic has already been used for robot navigation in cluttered environments [30, 31] and for person following application [32]. With the aim of having flexible and consistent robot behaviors, even when dealing with uncertainties, we use SpirOps AI, a middleware decision software dedicated to the design of autonomous decision processes. SpirOps software [28, 33, 36] provides a behavioral engine together with tools for easy design and debug of complex decisional models. A graphical editor (Figure 5) is used to edit independent Goals. Each Goal is a strategy producer to which the decision engine combines a Desire. Each Desire includes an Interest, that is the possible intention of strategy, a target and all the parameters required to define a complex strategy. The Interest can be seen as a fuzzy value, and the Goal as an extended 0+ fuzzy rule.

In order to define an appropriate behavior, the robot uses different strategies that may be selected dynamically according to the available information about the state of the person, the context and the environment. For the human following application we have defined five strategies: GoToUser, FollowUser, GlobalSearch, LocalSearch and UserFocus. The decision about the strategy to apply is made through fuzzy rules. These rules, which are defined by the programmer as expert systems, assign a degree of Interest to each strategy. The degree of Interest is computed as the product of different factors (person detection confidence and trajectory from EKF, interfaces inputs, obstacles, time elapsed from the last visit to a particular place, distance to the position to reach etc...) and it can take any value between 0 and 1. In the developed decision making engine, the computation of the Interests is left to SpirOps. More details about the computation model can be found in [37].

3.1.3. Strategies

The strategies defined to realize the user following behavior are briefly described here below:
GoToUser. This strategy is activated when the decision making engine determines that the robot may approach the person. This decision is taken when the robot has sufficiently high confidence on the person detection and location. The approaching distance is parametrizable. It can be set according to literature issues in proxemics [38] or to user preferences. The posture of the approached person is not considered for this strategy.

FollowUser. This strategy is activated when the decision making engine determines that the robot may follow the person. This decision is taken when the robot has sufficiently high confidence on the person detection and location and when it detects that the person is moving.

GlobalSearch. When the robot has to communicate with the person from an initial configuration which is far from him/her, it has first to search and localize him/her in the global environment. The criteria used to select the locations to explore among a predefined set of points is based on the time information of the last visit: the chosen point is located to the place which has been visited less recently. An example of rule associated for the GlobalSearch strategy is shown in Figure 6.

LocalSearch. This strategy is triggered when the robot fails to detect the target person. It consists in a local search at the most probable location of the person (estimated by the EKF). FocusUser. It is activated when the person is detected not moving or making small movements on the spot while he/she is communicating with the robot. The robot places itself in front of the person, where the environment allows this, and oriented towards him/her (the posture of the person is estimated from the face detector results).

For the robot paths and obstacle avoidance we rely on a path planner algorithm provided by the SLAM-based navigation integrated on the robot. We use a path following algorithm by using the control law proposed by Morin and Samson in [39].
3.2. Experiments

The human detection and the tracking behavior have been tested in a realistic domestic environment cluttered with chairs, tables, shelves and others obstacles of various kinds (Figures 7, 8 and 9). The goal of the robot is to keep a reliable detection of the first sensed person, so other people in the environment were considered as mobile obstacles. The 2D map of the environment is known by the robot through a previous exploration step (functionality provided by the SLAM).

3.2.1. Exiting through a door

The first experiment reproduces a situation very common in a real life scenario: a person exiting from a room. Since the robot has to keep him/her in its field of view, it must move according to the person’s trajectory. In this kind of situation, the robot behavior that we obtain by using our framework is shown in Figure 7. When the person passes through the door he/she is occulted by the wall. The decisional engine selects the LocalSearch strategy to make the robot search for the person in the position predicted by the EKF and the robot finds the person again.

3.2.2. Keeping detection in cluttered environment

The second experiment demonstrates the ability of the proposed framework to select the robot strategies by analyzing the context and the specific situation (environmental constraints, person detection confidence, person’s and robot state). At the beginning of the experiment, the person is not moving, the robot applies the GoToUser strategy and starts moving towards \(p_1\) (Figure 8). Then, the person starts walking and passes between two obstacles; the robot starts following him/her towards \(p_2\). For a little while the robot loses the person and selects the LocalSearch strategy, then the person is detected again and the robot applies the UserFollow strategy towards the point \(p_3\). The space between the two obstacles is not wide enough for the robot, therefore, the trajectory computed to reach \(p_3\) passes around an obstacle. Rapidly the person is lost again and the LocalSearch strategy is applied. In a bit the robot finds the person and selects the FollowUser strategy towards the point \(p_4\).

3.2.3. Face-to-face communication

A relevant social scenario for personal robots is the face-to-face communication with humans. The cognitive communication can take place through different kinds of interfaces: graphical, vocal, gestural, and so on. The quality of communication depends on how the devices (screen, microphone, camera etc.) are positioned with respect to the person. The aim of the last presented scenario is to show the benefit of the multimodal detector and of the decisional engine in cases of close communication. The robot exploits the multimodal detector to perceive the user and particularly the face detector to estimate its orientation. Having patrolled to find the person (GlobalSearch strategy), the FocusUser strategy is activated and the robot can place itself in front of the person and continue to focus on him/her throughout the face-to-face communication. Thereby the robot functionalities result easily accessible and the communication with the humans is more effective.

4. Physical based motion analysis

One of the potential uses of interactive mobile robots is the analysis of the motor activity of people in everyday tasks or during cognitive/physical exercises. Through their interactive capabilities of monitoring people and appropriate onboard sensors, they provide an answer to the problem of motion capture in natural environments. We focus here on human locomotor activity, a particularly complex motor function which can be affected by many factors leading to musculoskeletal or neurologic disorders (including chronic ankle sprains and low back pain, scoliosis, paroxysmal positional vertigo, cerebellar disease, Parkinson’s disease, vestibular deficits). The latter bring about some degenerations in the balance control which increase risks of fall. Analysis of locomotor activity is generally based on the spatio-temporal parameters (the Center of Mass (CoM) trajectory, the step length, the step frequency, the step width, the walking speed, the trunk acceleration etc.) that are considered to be relevant for gait evaluation [40, 44], but also on the parameters characterizing the state of dynamic balance of the postural system. Typically we use the Centre of Pressure (COP) which is the point where the Ground Reaction Force (GRF) vector intersects the plane of Base Of Support (BOS) of the feet on the ground. This reaction force represents the resultant of the contact force distribution on the ground.
In the last two decades, many technical solutions were developed to access these different parameters of human movement, e.g. treadmills, instrumented surfaces such as force platforms or particular shoes [41, 42]. To capture the kinematics movement, the most reliable results were achieved by systems such as Motion Capture (MoCap) which exploits high speed digital cameras to capture the 3D motion performed by a subject. Several systems can be used but most commonly the subject is fitted with either passive or active markers [43]. The passive markers are often covered with an infrared reflective material and then attached to the subject on predefined anatomical landmarks. The estimation of 3D human motions has been recently considered by using single or multi video cameras [45, 46]. The collected data is used to animate a model reproducing the anthropometric characteristics of the subject. The kinematic solver is formulated as a frame-by-frame weighted least-squares problem that minimizes the differences between the measured marker locations and the model’s virtual marker locations. Based on thus obtained joint motions, the dynamic results from the multi-body motion by exploiting a forward dynamic model [47]. These technologies are designed to be used in controlled laboratory conditions (with specific environmental set-up or instrumented clothes worn by the person) and cannot be employed for activity assessment in real life environments. Gonzáles et al. [48] proposed a non-invasive (marker-less) portable system based on the combination of 2 sensors, a Wii balance board and a Kinect to compute the CoM. This system is unsuitable for the analysis of locomotion because of the use of the balance board.

During locomotion, the most challenging phases for postural balancing system are beginning and end the walk, turning, avoiding obstacles (changing the length of the pitch, changing direction, stepping over objects, etc.) and bumping into people and objects. A system able to quantify human responses to perturbations induced by the daily life environment can have a use-value. To our knowledge there is no system that can be used for an in-depth analysis of human activity (especially of posture balance and stability) in non controlled environmental conditions, in a non-invasive fashion (marker-less method) and with an accuracy comparable with marker based technologies. We approach this need with a robot which can perceive and track humans while moving and with the embedded Kinect sensor to capture in real-time multi-segmental human motion by using the Microsoft Kinect SDK. The Kinect sensor can provide accuracy of depth reconstruction in the order of 1 – 4 cm in the range of 1 – 4 m but the estimation of the body geometry (limbs length) varies at each frame [49]. This makes the inference of others anthropomorphic data (mass, center of mass, inertia) based on geometric properties and the access to spatio-temporal and dynamic gait parameters difficult or impossible. We propose a method for an accurate and robust estimation of the anthropometric data from the Kinect skeletal detection as well as a physical based human model for gait and dynamic postural balance characterization.
4.1. Kinect Sensor

The Kinect sensor belongs to the class of devices known as depth cameras. These cameras interpret 3D scene information based on projected infrared light system called light coding that generates a 3D point cloud. The Kinect sensor has an approximate resolution of 1.3 mm per pixel and works at 30 Hz. The 3D point cloud is converted into a depth map and from a single depth map body parts are inferred using a random-forest based decision making approach, learned from over 1 million training examples. A mean shift algorithm is used to robustly compute 3D positions of body joints from the modes of discrete probability distributions. More details can be found in Shotton et al. [43].

The accuracy of the joints position provided by the Kinect skeletal detection algorithm has been investigated in [49] where the inter-joint distance (the limb length) variations have been estimated as more than 10 cm. For facing this significant inaccuracy, we propose a method for matching these data on a fixed limb-length skeleton of the human body after the initialization of the anthropometric data. The 3D virtual marker Cartesian positions thus obtained from this motion capture system are used to solve the joint trajectories of a virtual human model calibrated on anthropometric data from a physics based simulation.

4.2. Filtering method

Since serious errors exist in the estimated skeleton [49], an improved skeletonization solution to a space-time constraint problem is required. In previous works, the pose correction was approached in different ways: according to physical-model [50], by using Kalman-like filters [51], by adopting different kinds of regressors (random forest regressor [52], nearest neighborhood regressor [53], Gaussian process regression [54] etc.). We propose a method for pose correction that takes into account both the coherence of the anthropomorphic model of the skeleton and the spatial-temporal motion consistency. This method is composed of 3 steps (Figure 16).

Step 1: Initialization

When the person is in the Kinect field of view, his skeleton is detected and the 3D positions of all his joints are provided (Figure 10 left). During the initialization phase, we use a first set of Kinect measurements to define a reference model of the involved body. In particular, the system computes the anthropomorphic kinematic model of the body according to the mean distances between the consecutive joints.

Step 2: Physical model calibration

This step is required to keep the human physical skeleton consistent with the reference model (in anthropomorphic point of view). Our approach relies on a standard mathematical solution, constrained optimization. It consist of minimizing the scalar value quadratic objective function $f(X)$ while respecting a certain number of constraints $g(X)$ (Equation 7).

$$\text{minimize } f(X) \quad \text{subject to } g(X) = 0 \quad (7)$$

In the objective function $f(X)$ (Equation 8), $X$ is the vector containing all coordinates $(x_i, y_i, z_i)$ of dimension $(3 \times nJ)$ (3D coordinates for $nJ$ joints), such as $X = [x_0, y_0, z_0, \ldots, x_j, y_j, z_j, \ldots, x_nJ, y_nJ, z_nJ]$ and $a_i$, $b_i$, $c_i$ are the weights associated with the coordinates of the joint $i$ (in our case we consider equal weights on the 3 dimensions).

$$f(X) = \sum_{i=1}^{nJ} a_i (x_i - \bar{x})^2 + b_i (y_i - \bar{y})^2 + c_i (z_i - \bar{z})^2 \quad (8)$$

$$(x_i, y_i, z_i)$ and $(\bar{x}, \bar{y}, \bar{z})$ are respectively the coordinates of the joint $i$ provided by the skeleton detector algorithm and the coordinates that we are searching for to minimize the objective function while respecting the constraints in Equation 9.

$$g_j(X) = d^2(j_1, j_2) - d^2(M_j, M_{j+1}) = (x_j - x_{j+1})^2 + (y_j - y_{j+1})^2 + (z_j - z_{j+1})^2 - d^2_{j,j+1} = 0 \quad (9)$$

with $j = 1, \ldots, nJ - 1$, $(j_j, j_{j+1})$ consecutive detected joints, $(M_j, M_{j+1})$ consecutive joints in the model, and $d_{j,j+1}$ the distance between $M_j$ and $M_{j+1}$.

All the constraints are equality constraints and are fixed according to the reference model (initialization step). The optimization of the objective function is made by using the quadratic interior point method (QIPM), which is based on the improvement of initial conditions (measurements) for solving quadratic programming problems.

Step 3: Model Simulation and Parameters Analysis

Finally, the 3D trajectories of the kinetic skeleton joints virtual markers are used to animate a physical model of the subject on a dynamic simulator Arboris-Python [56]. Arboris-Python is an open-source constrained multibody dynamics simulation software written in Python language. It includes a generic and easily extensible set of joints (singularly-free multi-degree of freedom joints, non-holonomic joints, etc.) for the design and modeling of tree structure mechanisms with a minimal set of state variables. It gives access to the completed mechanical properties of the system as well as to the constraints and to the controllers implemented to get the desired behavior of the virtual human. Various control algorithms have been implemented in the Arboris-Python software, from the proportional-integral-derivative controllers to the predictive model based controllers which are used for the control of the locomotion and postural balance task or the interaction tasks with adaptive impedance. The equations of motion of these multibody systems are obtained with the Boltzmann-Hamel formalism [55] from which the first-order approximation of the model is computed. The resulting equations are then integrated using a time-stepping method and a semi-implicit Euler integration scheme. In this way, it is possible to introduce and solve additional constraints, i.e. the kinematic loops, which can be either unilateral (contact) bilateral (joint), with the help of a Gauss-Seidel algorithm [57]. For a complete description of this software you may refer to [58]. Once the model of the subject generated from a generic virtual human model (with 36 degrees of freedom) instantiated with the kinematic data retrieved from the calibration phase and by inference the anthropometric table Leva [58], the simulation is run using the 3D cartesian points as target joints for the selected joint axis through PID controllers. The input torque vector producing the virtual human motion in accordance with the one tracked by the Kinect are computed by solving a Linear Quadratic Program (LQP) that optimizes a set of weighted tasks (virtual joint marker trajectories and postural control) subject to equality and inequality constraints translating the physical limitations to implicitly satisfy the human motion. For the LQP problem formulation we refer to [59]. A number of parameters characterizing walking or steady state of the person can be obtained from the simulation of the virtual human (spatio-temporal gait parameters, energy, postural balance etc.). In the experiments presented hereafter we consider more particularly the CoP which is, by definition a point identical to the Zero Moment Point (ZMP). This is the point on the ground where the tipping moment acting on the biped, due to gravity and inertia forces, equals zero. The tipping moment is defined as the component of the moment that acting on the biped, due to gravity and inertia forces, equals zero. This method is required to keep the human physical skeleton consistent with the reference model (initialization step). The optimization of the objective function is made by using the quadratic interior point method (QIPM), which is based on the improvement of initial conditions (measurements) for solving quadratic programming problems.

$$f(X) = \sum_{i=1}^{nJ} a_i (x_i - \bar{x})^2 + b_i (y_i - \bar{y})^2 + c_i (z_i - \bar{z})^2 \quad (8)$$
as ankles and knees trajectories are proven be distinguishing factors useful to characterize the postural balance and to detect deviations during walk.

4.3. System validation

4.3.1. Material and procedure

In order to assess the reliability of the system, a preliminary set of experiments were made in a laboratory setting (Fig. 10 right). Five healthy subjects (see Table 2 for physical body data) were asked to execute 3 different movements (arm movement, rocking movement and side steps) on a posturography platform while wearing 13 CodaMotion markers, to validate the consistency of our system (see Fig. 10 center for marker placement). The Kinect embedded in the robot was simultaneously used for skeleton detection. The goal of this set of experiments was to assess the CoP trajectory provided by Arboris by comparing it with the CoP trajectory measured with a posturography platform (ground truth). The CoP was chosen because it's the only parameter that can be directly measured by a posturography platform. Both CodaMotion and Kinect data are collected and replayed by dynamic simulation and the resulted CoP trajectory was recorded and compared with the posturography platform measurements. All data was filtered using a second order lowpass Butterworth filter with cut-off frequency of 10 Hz. For Kinect data, the pose correction method described in Section 4.2 was applied.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Height (m)</th>
<th>Weight (kg)</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1.66</td>
<td>60</td>
<td>F</td>
</tr>
<tr>
<td>P2</td>
<td>1.78</td>
<td>78</td>
<td>M</td>
</tr>
<tr>
<td>P3</td>
<td>1.8</td>
<td>71</td>
<td>M</td>
</tr>
<tr>
<td>P4</td>
<td>1.75</td>
<td>88</td>
<td>M</td>
</tr>
<tr>
<td>P5</td>
<td>1.86</td>
<td>79.5</td>
<td>M</td>
</tr>
</tbody>
</table>

4.3.2. Results

To evaluate the consistency of our system with respect to the ground truth, the CoP trajectories computed by Arboris replaying both the Kinect and the CodaMotion data, and the CoP trajectories measured by the posturography platform were compared. In Figure 11, the 3 CoP trajectories related to arm movement executed by one subject are shown. Figures 12 and 13 report the CoP trajectories corresponding to rocking movement and to lateral steps for the same subject. The mean errors between these 3 trajectories were computed both for each singularly assessed movement (see Figures 18 and 19) and for each subject (Figures 14, 15 and 17). The differences in the trajectories were provided by Arboris by fitting on the mannequin first the Kinect and then the CodaMotion data (max mean error = 0.07 m), were obviously due to the approximative correspondence between the skeleton joints and the Codamotion markers (for imprecise markers placement). The max mean error between the trajectories provided by Arboris by using only the Kinect data (with the correction pose algorithm) and the ground truth was about 0.08 m with a standard deviation of ±0.03 m. These results support the idea that it is possible to estimate some important parameters related to human activity by using our system with a satisfactory accuracy. Indeed, not only the overall motion was well preserved during the simulation, but also the posture assessment showed good results.

4.4. Walking analysis during following

4.4.1. Material and procedure

System capabilities to extract some walking spatio-temporal and dynamics parameters of a subject have been evaluated in a real life situation. For this, the robot followed a walking person and tracked him by using the on board sensors under optimal conditions. The experiments were performed with a subject walking normally then the mobility of the lower limb joint was artificially manipulated in a controlled way. The objective of these experiments was to evaluate the sensitivity of the system to the disturbances, aiming the analysis of pathological walking activities later. Three experiments were conducted to produce permanent changes by mechanical effects on lower limb:
1. Subject walked in his comfortable walking speed (CWS) wearing his usual shoes;

2. Subject walked in his CWS wearing a brace on the right leg;

3. Subject walked in his CWS wearing a ski boot on the right foot.

Experiments 1 was the reference for the normal gait to which the results of the other experiments are compared. Experiments 2 and 3 aimed to evaluate the system accuracy in detecting anomalies in the
case of loss of mobility of the knee and ankle joints, respectively. The experiments were performed in the subject’s CWS to avoid the influence of velocity constraints in the deviations effects. The mechanical disturbances on the subject are shown in Figure 20.

The subject's movements were sensed by using only the Kinect embedded on the robot and running the skeleton detector algorithm provided by the Microsoft Kinect SDK. The collected data were then processed following the filtering method described in Section 4.2.

4.4.2. Results

The ZMP/CoP trajectory during walking is an important parameter for postural balance assessment. In Figure 21 the ZMP/CoP trajectories and the feet position during the walk are shown. The ZMP/CoP trajectory is computing by using the Linear Inverted Pendulum Model (LIPM). As expected, an useful factor to label walking deviations is the knees flexion in the support phase (Figure 22). When the subject walked wearing his usual shoes, the average flexion of right and left knees were almost equal (−0.24 rad). Wearing the brace in the right leg, the mean flexion of the right knee was lesser than the mean flexion of the left knee (−0.17 rad of right knee versus −0.24 rad of the left). The use of the ski boot entailed flexion difference (−0.23 rad of right knee versus −0.21 rad of the left).

The 2 types of lower limb disturbances tested are distinguishable observing the lift and side movements of the ankles, e.g. the movements orthogonal to the ground and to the CoM of the ankle in the swing phase. As shown in Figure 23, in the case of normal gait, both movements of right and left ankles were equal. The use of the brace entailed an asymmetry of side movement. The mean amplitude of the lateral movement observed for the right ankle was bigger than the mean value observed for the left ankle (0.07 m versus 0.12 m). The same phenomenon was observed when the subject walked wearing the ski boot (0.07 m versus 0.13 m). In this latter case, a slight asymmetry was observed for average ankles lift movement as well (0.11 m versus 0.13 m).

5. Conclusion

This paper proposes a system for the analysis of human activity in everyday life contexts using an autonomous robotic system. First of all, the framework to design interactive behaviors, particularly the person following behavior in cluttered environments is described and tested in different social relevant scenarios. This framework exploits the infor-
For all tested movements: CoP error on X between the trajectories provided by Arboris replaying Kinect data and CodaMotion (left); between the posturography platform measurements and the Arboris results with Kinect data as input (center); between the posturography platform measurements and the Arboris results with CodaMotion data as input (right).

For all tested movements: CoP error on Y between the trajectories provided by Arboris replaying Kinect data and CodaMotion (left); between the posturography platform measurements and the Arboris results with Kinect data as input (center); between the posturography platform measurements and the Arboris results with CodaMotion data as input (right).

Mechanical disturbances on the subject: normal (left), right knee immobilization (center) and ski boot on right foot (right).

Inclusion of a larger number of subjects could help to identify, in a clearer manner, the impact of the different disturbances on the gait parameters. Additional tests are planned to further validate the proposed system, including more participants and more kind of disturbances in order to have even more relevant and detailed results.
Figure 21. ZMP/CoP, COM trajectories and feet position during the support phase: normal walking (left), right knee immobilization (center) and ski boot on right foot (right).

Figure 22. Knees angular positions: normal walking (left), right knee immobilization (center) and ski boot on right foot (right).

Figure 23. Ankles lift (top) and side (bottom) trajectories: normal walking (left), right knee immobilization (center) and ski boot on right foot (right).
Acknowledgements

This research is partially supported by the project ANIPEV funded by European Regional Development Fund (ERDF) and the e-CareBot project funded the DREAM-IT Foundation.

References

[34] H. Hicheur, S. Vielle dent, M. J. E. Richardson, T. Flash and A. Berthoz. Velocity and Curvature in Human Locomotion along


