Fall Detection from Depth Map Video Sequences

Caroline Rougier[1], Edouard Auvinet[1], Jacqueline Rousseau[2], Max Mignotte[1], and Jean Meunier[1]

[1] Department of Computer Science and Operations Research (DIRO), University of Montreal, QC, Canada
{rougierc,auvinet,mignotte,meunier}@iro.umontreal.ca
[2] Research Center of the Geriatric Institute, University of Montreal, QC, Canada
jacqueline.rousseau@umontreal.ca

Abstract. Falls are one of the major risks for seniors living alone at home. Computer vision systems, which do not require to wear sensors, offer a new and promising solution for fall detection. In this work, an occlusion robust method is presented based on two features: human centroid height relative to the ground and body velocity. Indeed, the first feature is an efficient solution to detect falls as the vast majority of falls ends on the ground or near the ground. However, this method can fail if the end of the fall is completely occluded behind furniture. Fortunately, these cases can be managed by using the 3D person velocity computed just before the occlusion.

Keywords: fall detection, video surveillance, computer vision, 3D, depth map.

1 Introduction

1.1 Fall Detection Systems

Automatically detecting falls at home has become a major interest in research in these recent years. Indeed, the growing population of seniors in western countries motivates the development of new healthcare systems to ensure the safety of elderly people at home. In particular, for fall detection, different approaches have been explored with three main orientations. The first one is to place wearable sensors on the subject and detect falls with acceleration or rotation information. A detailed survey of this type of methodology is proposed by Noury et *al.* [7]. But these kinds of measure are intrusive in the subject's daily life. External (not wearable) sensors such as floor vibration detectors [1] could be another promising solution in the future but they require a complex setup and are still in their infancy. The other way to detect falls is to use a camera system with computer vision algorithms. For instance, monocular 2D methods were used to analyze the bounding box ratio of the person [11] or the 2D person velocity [6,10]. However, a problem with 2D velocity is that it is higher when the person is near the camera, so that thresholds to discriminate falls from a person sitting down abruptly, for

B. Abdulrazak et al. (Eds.): ICOST 2011, LNCS 6719, pp. 121–128, 2011.

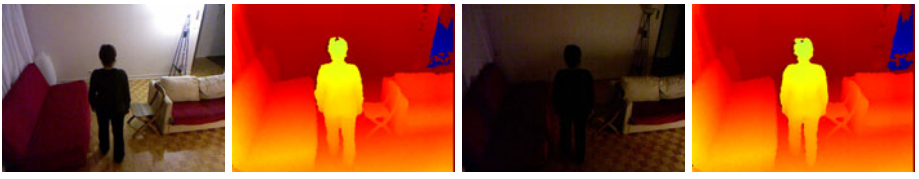
instance, can be difficult to define. This problem was solved using 2D information coupled with calibration data in order to deal with 3D real world coordinates. In this way, Rougier et *al.* [9] obtained the real 3D velocity vector with ellipsoidal head fitting. Another drawback of monocular systems is the management of occlusions. With 3D vision systems, the problem of detecting a falling person occluded by furniture becomes easier to solve. One solution is to reconstruct the total volume information in the scene, for example with a visual hull [2,3], but this is expensive and difficult to set up (multiple calibrated and synchronized cameras are needed). Other works use partial volume information [5] obtained from a Time-of-Flight sensor which returns precise depth images.

1.2 Depth Information

This depth information can be very useful for fall detection as it becomes possible to precisely track the person in the room. A depth image can be obtained with different methods:

- **Stereo vision** [16]. From two views of a scene, a depth image can be reconstructed. However, this type of system needs to be well calibrated and can fail when the scene is not sufficiently textured. Moreover, algorithms for stereo reconstruction are often computationally expensive. Notice that the usual stereo vision system cannot work in low light conditions. In this case, infrared (IR) lights can be added to the system but then, the color information is lost which generates segmentation and matching difficulties.
- **Time-of-Flight (TOF) camera** [17]. A TOF camera provides more accurate depth images than a stereovision system, but it is very expensive and currently limited to low image resolution (e.g. image size of 176x144 pixels in [5]).
- **Structured light**. A depth image cannot be obtained from a video sensor alone, but if a known artificial texture is added to the scene, a depth map can be recovered. This principle is used in the Kinect sensor [15] where an infrared structured light (IR dots) is projected in the scene and observed with an infrared camera. Such systems can acquire bigger images than a TOF camera at a lower price. For example, the Kinect sensor can acquire images with a size of 640x480 pixels at 30 fps, with a cost fifty times cheaper than a TOF camera. The drawback of this system is that depth information is not always well estimated at the boundary of objects and for areas too far from the IR projector.

To develop a low cost and easy-to-install fall detection system, the Kinect sensor [15] is a good solution to obtain depth images. An important advantage is that, with a depth image, the privacy of the person is readily preserved. Moreover, this solution can work day and night because of the use of an infrared sensor. Fig. 1 compares the usual video image with the depth map produced by the Kinect.



With normal light conditions **With low light conditions**

Fig. 1. Comparison of a video image with the corresponding depth map under different light conditions. The Kinect sensor works both day and night thanks to the IR sensor. The depth map blue areas are unreliable (e.g. too far from the projector or some object boundaries).

1.3 Our Method

The centroid height of the person relative to the ground is an efficient method for fall detection [5]. By pursuing this idea, we will simplify this method with several improvements and will demonstrate that a low cost depth map system can efficiently detect falls using this information. While other works need to localize the camera relative to the ground, we will show here that this information is not necessary. Moreover, unlike previous works, we manage the problem of occlusions by using the centroid velocity of the person to detect occluded falls, and we propose to use a training data set to learn the optimal detection thresholds. The different steps of our fall detection system are:

- **Ground plane detection** First, we automatically detect the ground plane of the room using the V-disparity approach which is explained in Section 2.
- **Person tracking and localization** The person is segmented from a depth background image and tracked to recover his/her 3D centroid localization. This step is described in Section 3.
- **Fall detection** The person 3D trajectory is analyzed to discriminate falls from normal activities as described in Section 4. More precisely, the human centroid height relative to the ground is used to detect (not or partially occluded) falls. When the end of an action is totally occluded by furniture, an analysis of the 3D body velocity prior to occlusion allows to detect the fall. Experimental results are shown in Section 5.

2 Ground Plane Detection

To compute the distance between the body centroid and the ground plane, we first need to automatically detect the ground plane in the scene. Once detected, the equation of the ground plane can be recovered and used to compute the distance between a 3D point of the scene and the ground plane.

The RANSAC plane fitting [12] is a commonly used method to fit a plane in the 3D space, but is rather computationally expensive. A recent method, called the V-disparity image [8,13], allows to detect the ground plane more easily with the depth image. Concretely, the V-disparity image consists in computing the

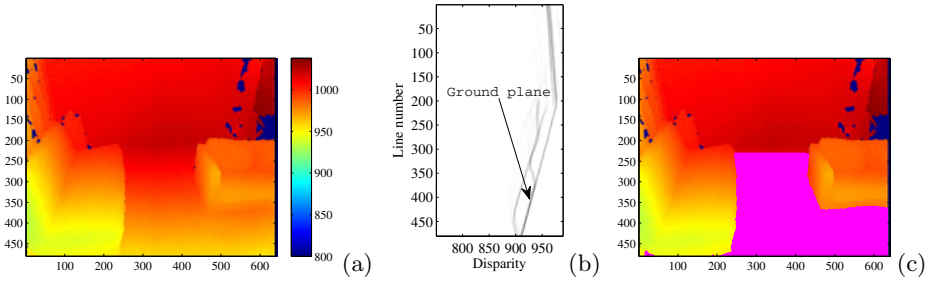


Fig. 2. Ground plane detection: (a) depth image, (b) V-disparity image and (c) ground plane segmentation (in magenta)

histogram of disparity values for each row of the depth image. An example of a V-disparity image is shown in Fig. 2(b). The straight line corresponding to the ground plane can then be extracted using the Hough transform, assuming that the floor represents a sufficiently large part of the scene.

With the V-disparity image and the line corresponding to the ground plane, floor pixels can be detected as shown in Fig. 2(c). From these pixels and their known depths, the 3D plane equation $ax + by + cz + d = 0$ of the ground plane can be recovered. The parameters a, b, c and d can be computed using a least squares fit of the 3D detected points.

3 Person Segmentation and Localization

A commonly used segmentation method for color and gray images [4] is used here to segment moving objects from a background image using depth images. A depth background image B is obtained from N_{train} background images ($N_{train} = 30$ for our experiments). The mean value and standard deviation are computed for each pixel of the image, and used for segmentation. For each pixel (i, j) of the current image I, the pixel is considered as foreground if $|I(i, j) - B(i, j)| \geq T(i, j)$ with the threshold $T(i, j)$ equal to 2 times the pixel standard deviation. Finally, the foreground image is cleaned with morphological filtering and the depth silhouette can be obtained by combining the depth image with the foreground silhouette. The 2D silhouette centroid (x_c, y_c) and the mean silhouette depth d_{mean} are computed respectively from the foreground silhouette and from the depth silhouette. The internal calibration parameters of the Kinect sensor [14], i.e. the focal length of the camera $(f_x, f_y) = (594.2, 591)$ and the coordinates of the principal point $(x_0, y_0) = (339.3, 242.7)$ in pixels, are used to obtain the 3D person localization relative to the camera coordinate system. The 3D silhouette centroid (X_C, Y_C, Z_C) is then obtained by:

$$\begin{aligned} X_C &= (x_c - x_0).d_{mean}/f_x \\ Y_C &= (y_c - y_0).d_{mean}/f_y \\ Z_C &= d_{mean} \end{aligned} \tag{1}$$

4 Fall Detection

From a video sequence, the person 3D trajectory is obtained in the camera coordinate system and analyzed to discriminate falls from normal activities. Falls can be detected during the post-fall phase when the person is motionless on the ground just after the fall [7]. The centroid height of the person relative to the ground has been used to detect an abnormal position near the floor [5]. This method works well when the silhouette is not occluded by furniture in the room. In [5], the camera localization relative to the ground was needed. Here this information is not necessary, since we just need to compute the distance from a point (body centroid) to a plane (floor).

4.1 3D Distance From the Ground Plane

The distance from the 3D centroid to the ground plane can directly be obtained by a simple point-plane distance:

$$D = \frac{|aX_c + bY_c + cZ_c + d|}{\sqrt{a^2 + b^2 + c^2}} \tag{2}$$

The distance D can be directly used to check the location of the body relative to the ground when the person is not occluded. However, in case of total occlusion at the end of the fall, this distance can not be computed. As we use only one Kinect per room, an occlusion can happen because of furniture (e.g. sofa) in the scene. In this case, we use another characteristic, the 3D body velocity, to analyze what happened just before the occlusion.

4.2 3D Body Velocity

Another weakness of previous works [5] is that they did not deal with occlusions which is a difficult problem in video surveillance. In this work, a special analysis is done when an occlusion occurs. Indeed, the method based on the height relative to the ground can fail when the person is completely hidden by furniture of the scene like a sofa. Our idea here is to analyze the body velocity, just before the occlusion occurs, to try to guess what happened. The body velocity V was computed as the centroid displacement over a one second period. Generally, V will be lower when a person is doing normal activities than when a person falls. Notice that this criterion is only used in case of occlusion, because in other cases, it can be prone to many false positives, for example when the person brutally sits down in the sofa.

4.3 Automatic Fall Detection

Unlike previous works [5], we investigate an automatic method for fall detection, with a training data set to determine the best thresholds. The training data set was composed of normal activities like walking, sitting down and crouching down, with some occluded activities. The centroid distance from the ground

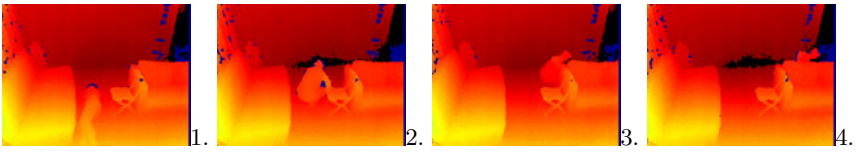
D_{train} and the 3D body velocity V_{train} were computed in the video sequence. Then, they were used to automatically define two thresholds from the mean value and standard deviation with a 97.5% confidence interval:

$$T_{Dmin} = \overline{D}_{train} - 1.96\ \sigma_{D_{train}} \qquad T_{Vmax} = \overline{V}_{train} + 1.96\ \sigma_{V_{train}} \tag{3}$$

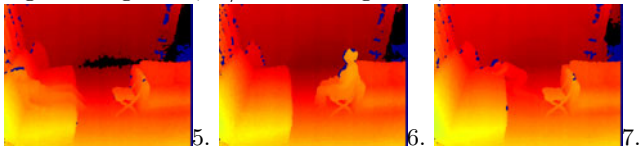
If the body distance from the ground D is lower than T_{Dmin}, then a fall is directly detected. If an occlusion is detected (silhouette which suddenly completely disappears), we search for a high body velocity in the few frames (30 frames = 1 second in our experiments) before the occlusion higher than the threshold T_{Vmax}.

5 Experimental Results

Our fall detection system has been tested on simulated falls and normal activities (like walking, sitting down, crouching down) recorded with a Kinect sensor [15] for a total duration of 4 minutes 9 seconds. With this training data set, the computed detection thresholds were $T_{Dmin} = 35.8\ cm$ and $T_{Vmax} = 0.63\ m/s$. Our fall detection method was validated with a data set composed of 30 sitting down actions, 25 falls including 7 totally occluded, and 24 crouching down actions



1. fall ending on the ground, 2./3. crouching down, 4. occluded crouching down



5. sitting down in the sofa, 6. sitting down on a chair, 7. fall ending on the sofa

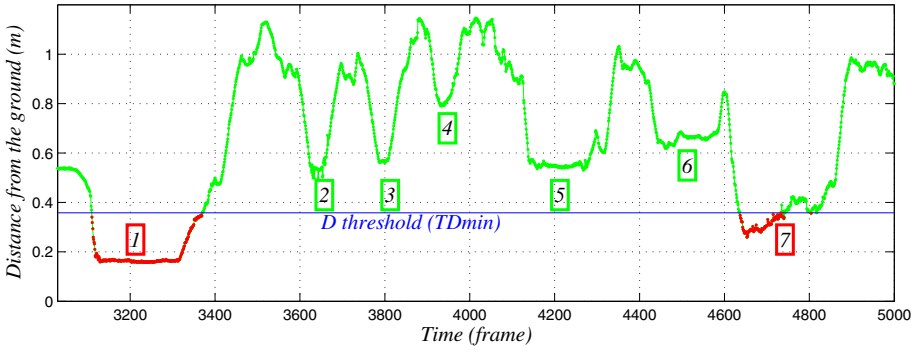


Fig. 3. Distance-from-the-ground curve (bottom) obtained for not or partially occluded events (numbered depth maps above)

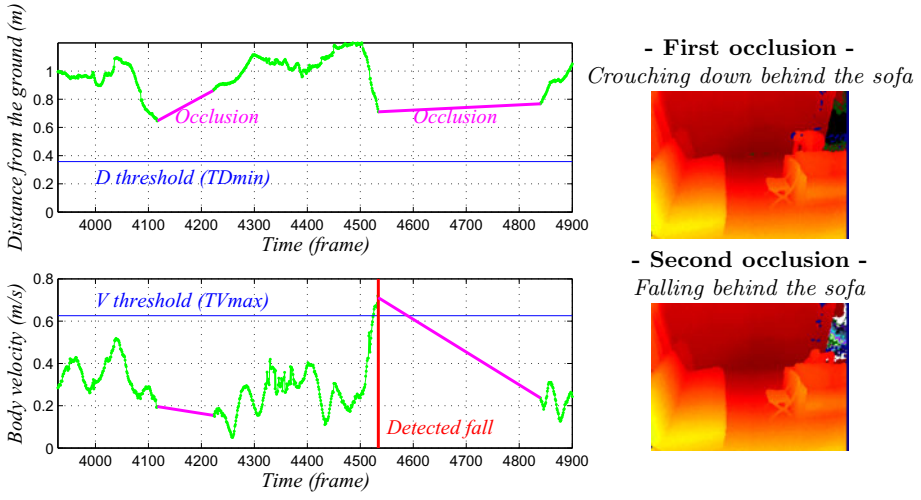


Fig. 4. Distance from the ground and body velocity curves (left) obtained for occluded events (right). The fall is correctly detected with a high body velocity.

including 6 totally occluded, for a total duration of 15 minutes 34 seconds. Some examples are shown in Fig. 3 and 4.

All 'not occluded' events were correctly classified by using the body distance from the ground plane. Even when a fall ended partially on the sofa, the fall was correctly detected as shown in Fig. 3. In case of total occlusion, the body velocity is an interesting method for fall detection, but can be difficult to use to discriminate a fall from a person who brutally sits down in a sofa. Therefore, this feature is only used in case of occlusion to discriminate a fall from a person who crouches down behind a sofa. An example of such cases is shown in Fig. 4. Only one fall was not detected because the person grabbed the sofa, which slowed down the fall, before finishing totally occluded by the sofa at the end. In this case, the body velocity was not sufficient to detect the fall properly.

6 Discussion and Conclusion

When a fall is not occluded, the height relative to the ground is an efficient feature for fall detection as most falls ends on the floor. However, for a system usable in real life, the occlusion problem must be addressed. With a special analysis for occluded events, our fall detection system is also able to detect falls ending totally occluded. Even with cheap depth sensors, our fall detection method gives really good detection results with an overall success rate of 98.7% in our experiments. This solution preserves the privacy of the elderly and can work day or night. For future work, we plan to upgrade the background to avoid 'ghosts' generated by moving objects (e.g. chair) during the segmentation step. Currently, our method is developed with Matlab® which do not provide real-time code, but could easily run in real-time with a C/C++ implementation.

Acknowledgments. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Alwan, M., Rajendran, P., Kell, S., Mack, D., Dalal, S., Wolfe, M., Felder, R.: A smart and passive floor-vibration based fall detector for elderly. 2nd Information and Communication Technologies 1, 1003–1007 (2006)
2. Anderson, D., Luke, R.H., Keller, J.M., Skubic, M., Rantz, M., Aud, M.: Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Computer Vision and Image Understanding 113(1), 80–89 (2009)
3. Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J., Meunier, J.: Fall detection with multiple cameras: An occlusion-resistant method based on 3D silhouette vertical distribution. IEEE Trans. on Information Technology in Biomedicine (2010)
4. Collins, R., Lipton, A., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O.: A System for Video Surveillance and Monitoring: VSAM Final Report. Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University (May 2000)
5. Jansen, B., Temmermans, F., Deklerck, R.: 3D human pose recognition for home monitoring of elderly. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Lyon, France, pp. 4049–4051 (2007)
6. Lee, T., Mihailidis, A.: An intelligent emergency response system: preliminary development and testing of automated fall detection. Journal of Telemedicine and Telecare 11(4), 194–198 (2005)
7. Noury, N., Fleury, A., Rumeau, P., Bourke, A., Laighin, G., Rialle, V., Lundy, J.: Fall detection - principles and methods. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 1663–1666 (2007)
8. Rebut, J., Toulminet, G., Bensrhair, A.: Road obstacles detection using a self-adaptive stereo vision sensor: A contribution to the ARCOS french project. In: IEEE Intelligent Vehicles Symposium, pp. 738–743 (2004)
9. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Robust Video Surveillance for Fall Detection based on Human Shape Deformation. IEEE Transactions on Circuits and Systems for Video Technology (2011)
10. Sixsmith, A., Johnson, N.: A smart sensor to detect the falls of the elderly. IEEE Pervasive Computing 3(2), 42–47 (2004)
11. Töreyin, B., Dedeoglu, Y., Çetin, A.: Hmm based falling person detection using both audio and video. In: Proc. IEEE International Workshop on Human-Computer Interaction (2005)
12. Yu, Q., Araujo, H., Wang, H.: A Stereovision Method for Obstacle Detection and Tracking in non-flat Urban Environments. Autonomous Robots 19(2), 141–157 (2005)
13. Zhao, J., Katupitiya, J., Ward, J.: Global Correlation Based Ground Plane Estimation Using V-Disparity Image. In: IEEE International Conference on Robotics and Automation, pp. 529–534 (2007)
14. Kinect calibration parameters, `http://nicolas.burrus.name/index.php/Research/KinectCalibration`
15. Kinect, 3D-sensing technology, `http://www.primesense.com`
16. Stereo cameras, Point Grey Research Inc., `http://www.ptgrey.com`
17. MESA Imaging, `http://www.mesa-imaging.ch/index.php`