Local Behavior Modeling based on Long-Term Tracking Data

Rainer Planinc and Martin Kampel
Vienna University of Technology
Favoritenstr. 9-11/183-2
A-1040 Vienna
rainer.planinc@tuwien.ac.at

Abstract

Modeling the behavior of elderly people to detect changes in their health status or mobility is challenging and thus requires to combine temporal and spatial knowledge. Spatial knowledge is obtained by a novel human centered scene understanding approach, being able to accurately model sitting and walking regions based on noisy long-term tracking data from a depth sensor, without exploiting geometric information. A local behavior model based on the detected functional regions is introduced, allowing an in depth behavioral analysis. The proposed approaches are evaluated on three different datasets from two application domains (home and office environment), containing more than 180 days of tracking data.

1 Introduction

The proposed behavior modeling approach introduces a spatio-temporal model, modeling the behavior of a person in the home environment: regions of interests are detected by a novel human-centric scene understanding approach and temporal behavior is modeled within parts of the scene locally. In contrast to traditional object-centered scene understanding (e.g. [1, 2, 3]), human centered scene understanding focus on functional aspects based on information how the person interacts with the scene and which functionality the scene offers for a human [4]. Please note that in the context of the proposed approach, the taxonomy of Chaaraoui et al. [5] is used, defining *behavior* to have a time frame of days and weeks and reflecting the way of living and personal habits.

The use of long-term tracking of humans in order to describe object functions within a room is introduced by Delaitre et al. [6]. Their approach combines pose analysis (standing, sitting and reaching) together with object appearance (geometrical information) to model the interaction between human actions and objects. Delaitre et al. [6] state that the use of pose estimation is still challenging, since the pose estimation does not work robustly when being applied in practice and introduces wrong pose estimations. Although Delaitre et al. [6] use long-term tracking to model person-object interactions, their analysis is based on time-lapse videos, only offering discrete but not continuous information (i.e. snapshots). In contrast, a human-centric scene modeling approach based on depth videos is introduced by Lu and Wang [7]. For this proof of concept, pose estimation is performed on depth data and objects are modeled as 3D boxes within the scene. Together with geometric knowledge (estimation of vanishing points) of the scene, a room hypotheses including supporting surfaces for human actions is created and walkable areas are estimated. However, analysis and evaluation of the algorithm is performed on only several minutes of data and only deals with a few depth frames, but not a long-term analysis. Moreover, the authors [7] state that skeleton data could theoretically be used as well, but is not stable enough to obtain reasonable results, since skeleton data is noisy and defective. In summary, long-term people tracking is either performed by the use of time-lapse videos [6], is based on still-image pose estimation [8] or is a proof of concept [7]. Hence, the introduced approach does not focus on snapshots, but uses the tracking data and pose information of humans being continuously tracked 24/7 with a depth sensor (Asus Xtion pro live). Thus results in detailed tracking data throughout the day and allows to model the scene according to the functions being used by the human, but also to model the behavior within the scene throughout the day.

Temporal aspects to model behavior are considered by Floeck & Litz [9] and Cuddihy et al. [10], where activity data is obtained from different sensors within a flat. Floeck & Litz [9] learn an inactivity profile from the sensor data in order to model the temporal behavior, but do not consider spatial aspects. Abnormal inactivity is detected by comparing the trained reference profile (e.g. average inactivity profile of one month) to the current inactivity profile. Similar to [9], Cuddihy et al. [10] use inactivity profiles. A reference alert line is trained within the last 45 days and an alarm is generated if he inactivity within a time slot raises the pre trained alert line. Planinc et al. [11] propose to use activity histogram comparisons rather than modeling inactivity. Instead of creating an inactivity profile, the activity is modeled using histograms. The authors of [11] suggest to use 24 bins for creating the histogram, i.e. one bin per hour. A reference activity histogram is learned during the training phase and histograms are compared on a per day basis with the trained reference histogram in order to detect deviations from the behavior. However, the approaches of [9, 10, 11] only consider temporal aspects since the location of activity is not integrated within these approaches.

In conclusion, current approaches either focus on temporal *or* spatial (object-centered) aspects of the scene. Hence, the combination of spatial and temporal knowledge is proposed and results in a solid foundation for a novel behavior model. Based on the findings of the related work, the aim of this paper is twofold: first, a novel approach of scene understanding using long-term tracking data obtained by a depth sensor is introduced. Thus allows to obtain functional regions within a scene, being used for different activities without incorporating geometrical information. Tracking data is obtained by tracking the human 24/7 and thus allows a detailed spatio-temporal analysis of the scene.

Hence, the second aim of this paper is to model the behavior of (elderly) people not globally, but locally within the semantic regions of the scene, i.e. the walking and sitting behavior is analyzed. Although this approach focus on the detection of health changes of elderly, it can be applied to different domains as well (e.g. indicators of stress at the workplace can be detected). The rest of this paper is structured as follows: Section 2 describes the proposed approach, an evaluation and results are presented in Section 3. Finally, a conclusion is drawn in Section 4.

2 Methodology

The proposed approach combines the advantages of 3D depth data together with long term tracking information and introduces a new local behavior model in order to detect health-related changes. Depth data is obtained by the use of an Asus Xtion pro live, the detection and tracking of the person is performed using the OpenNI SDK [12]. The 3D position of a person within a frame is obtained from long-term tracking data, where filtering mechanisms to reject unreliable tracking data are applied. The filtered long-term tracking data is clustered according to the height (distance to the ground floor) into walking and sitting clusters. Kernel density estimation together with non maxima suppression and the calculation of a convex hull yields in "hotspots" of each activity region. Finally, an behavioral model is learned locally within in each activity region.

2.1 Spatial Knowledge

Since tracking data is noisy, data need to be filtered in order to obtain reliable data. Filtering is based on the following three features: 1) it can be assumed that a person being either walking or sitting is in an upright pose (body orientation), 2) the tracked Center of Mass (CoM) is within a plausible range of height (i.e. lower than 1.6 m) and 3) the confidence values of OpenNI are used to eliminate "unreliable" CoM values in order to ensure correct tracking data. The filtering process is introduced in order to eliminate tracking errors, which are caused by objects being recognized as person and being wrongly tracked. The height of the CoM is clustered for each frame using the k-means algorithm, resulting in two cluster: sitting and walking. K-means is chosen since it is fast and the number of cluster is known in advance. A kernel density estimation with a bivariate Gaussian kernel is performed in order to estimate the probability density function of the clustered data. The estimation is performed on both classes (sitting and walking) separately to model the probability density function of the walking and the sitting class. This step is performed to detect "hotspots" within the clustered data, i.e. areas being relevant for this class. Non maxima suppression is applied to suppress irrelevant areas and allows to focus on the main areas, being representative for each class. The relevant area is obtained by thresholding the probability density function with a fixed threshold, where the resulting contour describes the functional areas within the scene. In order to aggregate smaller but similar areas, the convex hull of all contours is calculated to ensure a coherent area.

2.2 Temporal Knowledge

Temporal Information is obtained from an extended version of Planinc et al. [11], introducing a behavior model based on histograms. Thus allows to model the behavior on a global level, considering only temporal but not spatial information. By extending this approach to model behavior on a local level based on the obtained regions, a spatio-temporal approach is introduced. Hence, instead of calculating one global behavior histogram for the environment, a local behavior histogram is calculated for each detected region individually. Tracking information within the detected areas is aggregated and a reference histogram H^i_{ref} with 24 bins (one bin represents one hour) is trained for all regions i separately. The reference histogram H^i_{ref} is the average histogram of all n training days within the respective region.

The distance between the histogram to be trained or tested H^i_t and the reference histogram is calculated using the chi-square distance [13]

$$d^i_t(H^i_t, H^i_{ref}) = \frac{1}{2}\sum_k \frac{(H^i_t(k) - H^i_{ref}(k))^2}{H^i_t(k) + H^i_{ref}(k)} \quad (1)$$

The average distance $\overline{d}_i$ represents the average distance between all n training histograms and the reference histogram H^i_{ref}. In combination with the standard deviation σ_i, deviations from the reference histogram are detected if

$$|d^i_t| \geq \overline{d}_i + \sigma_i \quad (2)$$

In other words, if the distance between the reference histogram H^i_{ref} and the current histogram H^i_t exceeds the threshold, a deviation from normal behavior is detected. This allows to model the behavior within sitting and walking areas separately and thus allows to get insights about the walking and sitting behavior individually. Moreover, a change of walking and sitting behavior can be detected (e.g. less walking and increased sitting behavior due to the reduction of mobility), which is not possible with the use of a global behavior model since actions and activities can not be separated.

3 Results

The evaluation is split into two parts: the evaluation of the human centered scene understanding and the performance evaluation of the local behavior model. The evaluation is based on three different datasets: the monitoring of a kitchen, living room and office environment. In all scenes, a sitting area as well as a walking area is present and to be detected by the proposed approach. Although the focus is the monitoring of elderly people and thus home environments, the authors want to demonstrate the possibility to extend the proposed approaches to new environments as well (e.g. office). The advantages of monitoring elderly at home are obvious (detect health deterioration, increased or decreased mobility), but are more subtle in the office environment. However, also in the office environment, the health status can be detected automatically: longer working hours than usual and less breaks are examples

	Walking		Sitting	
Dataset	Unfiltered	Filtered	Unfiltered	Filtered
Kitchen	0,84	0,87	0,44	0,33
Living room	0,35	0,73	0,70	0,75
Office	0,90	0,88	0,65	0,68

Table 1. Resulting F-scores of the proposed scene understanding approach

to indicate stress. Hence, the system can be adopted in order to detect stress at the workplace.

The evaluation of the proposed approaches is performed on 90 days tracking data of the living room, 74 days of tracking data in the kitchen and 20 days of tracking inside the office. Evaluation of the behavior modeling approach is not performed on the office data set since it only contains 20 days of monitoring and thus does not contain enough data for a long-term behavior analysis. Hence, it is only performed on the living room and kitchen dataset. Previous experiments have shown that the system is sensitive and produces false alarms - thus the evaluation is based on the fact, that no changes of mobility are present and thus no alarms should be generated.

3.1 Scene Understanding

Table 1 shows the results of the proposed approach: the human centered scene understanding approach is evaluated under 3 different scenes, where the sitting and walking areas were annotated as ground truth in advance. The f-scores are calculated using the maximum number of training days and shows that the proposed filtering mechanism is able to improve the quality of the scene understanding approach. Moreover it is shown that the proposed approach is able to model the sitting and walking areas within different indoor scenes robustly. The proposed filtering does not dramatically change the results in the office dataset since it is not as challenging as the kitchen and living room training set - this is due to the fact that almost no tracking errors are present in the office dataset, but severe tracking errors occur in the living room and kitchen dataset (e.g. doors are tracked).

In order to perform further analysis on these quantitative results, qualitative evaluation of the results is shown in Figure 1: for the kitchen dataset, two different models are obtained (top) - the correct one depicted on the left side and a wrong one depicted on the right side. The wrong model is generated due to wrong tracking data and thus, the sitting area is modeled much bigger than it actually is. This strong influence of the wrong tracking data is a result of too less training data, hence when using more training data, the influence of each training day is minimized. However, the low f-score of the sitting area in the kitchen dataset is also a result of the ground truth labeling - as can be seen in the top left image of Figure 1, the detected sitting area considers only one bench while the other bench is ignored since the second bench is already outside the tracking range. Hence, ground truth annotation need to be adopted according to the range limits of the tracking algorithm. In contrast, the walking area in the kitchen dataset is modeled very well

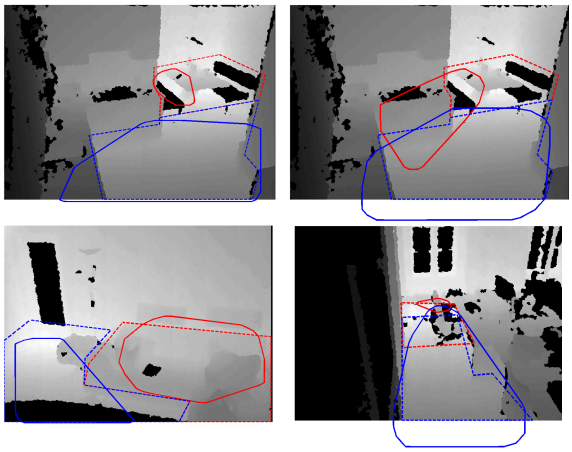

Figure 1. Results of the scene understanding approach: sitting (red) and walking areas (blue) in the kitchen (top), living room (bottom left) and office dataset (bottom right) with respective ground truth marked with dotted lines

and thus achieves a high f-score. Qualitative analysis of the living room and office dataset show that the regions are modeled accurately and confirm the quantitative results: Figure 1 (bottom) depicts the detected sitting (red) and walking area (blue) of the living room (left) and office dataset (right). The corresponding ground truth is shown with dotted lines. Similarly to the kitchen dataset, also in the office dataset, the ground truth of the sitting area is larger since all possible positions to sit are considered within the ground truth annotation - however, this does not mean that all possible positions to sit are actually used by the person, since people tend to usually sit on the same spots and do not change these spots often.

3.2 Region Based Behavior Modeling

The aim of region based behavior modeling is to gain a more accurate local behavior model, representing different activities based on the actions sitting and walking. In order to evaluate the performance of the proposed approach, the number of false alarms is used as indicator since during the testing period no true positives were obtained and thus a ROC (Receiver Operating Characteristic) curve can not be constructed. Moreover, since behavior models are prone to false alarms, the performance of the system in order to reduce the false alarms is evaluated. For repeated subsampling, 20 different training samples are chosen randomly and the evaluation is performed 20 times, where results are averaged. The results of the region based behavior modeling are depicted in Figure 2 and Figure 3[1]: although the proposed approach is able to outperform the approach of Cuddihy et al. [10] in terms of lower false alarms independently from the behavior model, the approach of Planinc et al. [11] is only outperformed when focusing on the sitting region. It can be seen that the walking area results in a higher number of false alarms than the global approach of Planinc

[1]Please note that the number of false alarms is cut off in order to focus on the details.

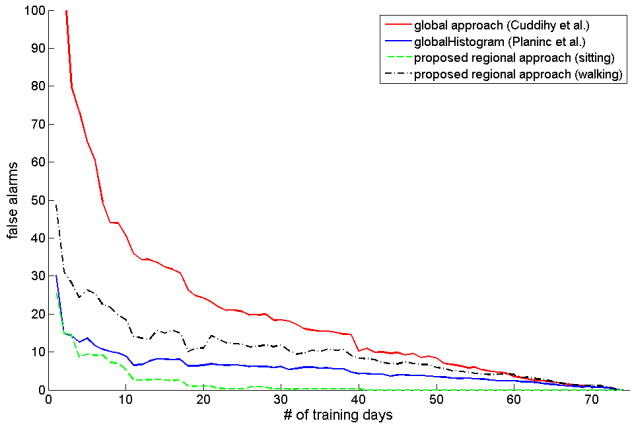

Figure 2. False alarm rate depending on the training (kitchen)

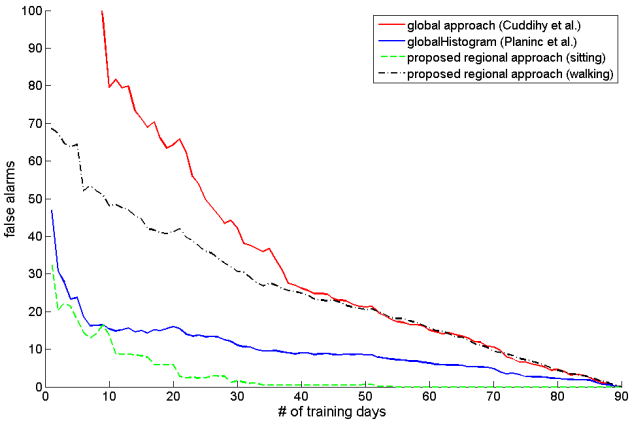

Figure 3. False alarm rate depending on the training (living room)

et al. [11], since the walking data is not as regularly organized as the sitting data is. This can be explained easily due to the fact, that people tend to sit at the same time every day, e.g. to eat lunch or dinner, or to watch TV in the evening. Hence, a more structured daily routine is found for the time we spend sitting, whereas walking is more unstructured since we usually do not walk to e.g. clean the flat at the same time but in a more unstructured way. Hence the proposed approach is able to incorporate these aspects of the daily routines, whereas this information is lost when using a global approach.

4 Conclusion

This paper introduced a novel method for human centered scene understanding, based on continuous tracking data obtained from a depth sensor. In contrast to other approaches, no geometric information of the scene is used in order to obtain functional areas within a scene. The focus on walking and sitting areas was chosen since these are the most popular functions within a room. Quantitative and qualitative evaluation showed that the proposed approach is able to accurately model sitting and walking areas based on the information, which regions people are using for walking and sitting. Moreover, the modeling of spatio-temporal behavior in order to detect health related changes of elderly people was introduced by modeling the behavior in each region locally and thus allowing an in-depth analysis of mobility. Since the application to other areas was shown exemplary by extending the approach to an office, future work deals with the extension to other application domains and a long-term study with elderly, since mobility does not change within days or weeks, but within months and years.

Acknowledgement. This work is supported by the EU and national funding organisations of EU member states (AAL 2013-6-063).

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[2] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images," in *CVPR*, 2013, pp. 564–571.

[3] J. Mutch and D. G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features," in *CVPR*, vol. 1, 2006, pp. 11–18.

[4] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert, "From 3D scene geometry to human workspace," in *CVPR*. IEEE, Jun. 2011, pp. 1961–1968.

[5] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 873–10 888, Sep. 2012.

[6] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros, "Scene semantics from long-term observation of people," in *ECCV*, Florence, 2012, pp. 284–298.

[7] J. Lu and G. Wang, "Human-centric indoor environment modeling from depth videos," in *ECCV-Workshops and Demonstrations*, 2012, pp. 42–51.

[8] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic, "People Watching: Human Actions as a Cue for Single-View Geometry," in *ECCV*. Springer Berlin Heidelberg, 2012, pp. 732–745.

[9] M. Floeck and L. Litz, "Activity- and Inactivity-Based Approaches to Analyze an Assisted Living Environment," in *SECURWARE*, 2008, pp. 311–316.

[10] P. Cuddihy, J. Weisenberg, C. Graichen, and M. Ganesh, "Algorithm to automatically detect abnormally long periods of inactivity in a home," in *ACM SIGMOBILE workshop on systems and networking support for healthcare and assisted living environments*. New York, NY, USA: ACM, 2007, pp. 89–94.

[11] R. Planinc and M. Kampel, "Detecting Unusual Inactivity by Introducing Activity Histogram Comparisons," in *VISAPP*. Lisbon, Portugal: SCITEPRESS, 2014, pp. 313–320.

[12] "OpenNI," http://www.openni.org, 2011, [Online; accessed 10-April-2014].

[13] S.-H. Cha, "Taxonomy of nominal type histogram distance measures," in *Proceedings of the American Conference on Applied Mathematics*, ser. MATH'08, 2008, pp. 325–330.