Human Centered Scene Understanding based on 3D Long-Term Tracking Data

Rainer Planinc and Martin Kampel

Computer Vision Lab, Vienna University of Technology
Favoritenstr. 9-11/183-2, A-1040 Vienna
rainer.planinc@tuwien.ac.at

Abstract. Scene understanding approaches are mainly based on geometric information, not considering the behavior of humans. The proposed approach introduces a novel human-centric scene understanding approach, based on long-term tracking information. Long-term tracking information is filtered, clustered and areas offering meaningful functionalities for humans are modeled using a kernel density estimation. This approach allows to model walking and sitting areas within an indoor scene without considering any geometric information. Thus, it solely uses continuous and noisy tracking data, acquired from a 3D sensor, monitoring the scene from a bird's eye view. The proposed approach is evaluated on three different datasets from two application domains (home and office environment), containing more than 180 days of tracking data.

Keywords: long-term tracking; human-centric; scene understanding;

1 Introduction

Traditional scene understanding is object centered (e.g. [3,6,9]) rather than human centered. Human centered scene understanding focuses on functional aspects based on information not provided by the scene and objects itself, but on information how persons interact with the scene and which functionality the scene offers for a human [5]. The use of long-term tracking of humans in order to describe object functions within a room is introduced by Delaitre et al. [2]. Due to the combination of pose analysis (standing, sitting and reaching) and object appearance (geometrical information), the interaction between human actions and objects are modeled. 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 [2]. However, based on performing long-term tracking, this effect can be minimized by enhancing the amount of tracking data and thus enhancing the accuracy of the pose estimation at a specific location. Although Delaitre et al. [2] use long-term tracking to model person-object interactions, their analysis is based on time-lapse videos, only offering discrete but not continuous information (snapshots).

While Delaitre et al. [2] recognize the objects, Fouhey et al. [4] extends their approach by not only recognizing objects, but modeling the scene in 3D, based on

the object functionality. Again, time-lapse videos are analyzed to detect the pose and combined with a 3D room geometry hypothesis in order to model the room based on the functionality it is offering. Similar to [2], the authors focuses on the poses standing, sitting and reaching in order to classify surfaces into walkable, sitable and reachable surfaces. Based on the pose information, estimates of the functional surfaces are generated and combined with the geometric information obtained by the room hypothesis.

In contrast, a human-centric scene modeling approach based on depth videos is introduced by Lu and Wang [8]. For this proof of concept, a background model is learned from the depth data and used to obtain the human silhouette. In combination with pose estimation performed on the silhouette of the person, 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 [8] 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.

Long-term human centered scene understanding is either performed by the use of time-lapse videos [2], is based on still-image pose estimation [4] or is a proof of concept [8]. Hence, the contribution of this paper is twofold: first, a novel human-centric scene understanding approach based on continuous long-term tracking data of humans is introduced. Second, the proposed scene understanding approach does not incorporate geometric information and thus is solely based on long-term tracking data (position and pose). With the proposed approach, a scene can be modeled according to the functions being used by the human. The proposed approach is evaluated during a long-term evaluation over the duration of more than 180 days of tracking data. The rest of this paper is structured as follows: Section 2 describes the proposed approach, an evaluation is 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 and introduces a novel approach for human-centric scene understanding, solely based on noisy tracking information. 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 [1][1]. 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

[1] Since Primesense is not supporting the OpenNI project any longer, the authors would like to stress that the proposed approach is fully independent from third party companies. Hence, other depth cameras and tracking algorithms can be used in order to obtain the 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, representing areas being commonly used by humans. In the following sections, detailed information about the proposed filtering, clustering and modeling of the regions are provided.

2.1 Filtering

Long-term tracking data contains noise and tracking errors (e.g. furniture is wrongly tracked as a person), influencing the results of the proposed approach. Hence, tracking data is filtered according to plausibility rules in order to ensure robust results. Although the filtering step removes a high amount of tracking information (i.e. all other activities and tracking information except walking and sitting), this does not influence the overall performance since the approach focuses on long-term tracking, thus ensuring a sufficient number of reliable information is available. 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. During the initial filtering step the skeleton data is used in order to detect the orientation of the body, based on the idea of Planinc and Kampel [10].

2.2 Clustering

Interesting functional areas within indoor environments are sitting and walking areas, and thus are in the focus of research (e.g. [5,2,4]). Hence, tracking data (CoM) is clustered into sitting and walking areas using the k-means algorithm. In a first step, the ground floor parameters are calculated using OpenNI and the height of the CoM for each tracking position is calculated. Tracking data is clustered for each frame according to its distance to the ground floor (=height), 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. However, this approach assumes that both, a sitting and walking area are within the field of view. If only one type of area is within the field of view (i.e. either sitting or walking), the obtained clusters need to be merged (e.g. by analyzing the distance of the cluster centers).

2.3 Kernel Density Estimation

A kernel density estimation with a bi-variate 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 in order to model the probability density function of both classes. 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 applying a fixed threshold to the probability density function, 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. Figure 1 illustrates an example of the proposed workflow: unfiltered tracking points of the sitting class (left) are filtered and a kernel density estimation as well as non-maxima suppression is performed (middle). The corresponding contour plot visualizes the sitting area (right).

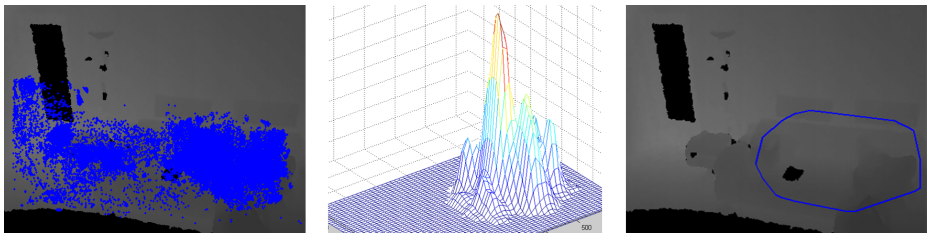

Fig. 1. Example of the workflow: unfiltered tracking data (left), kernel density estimation (middle) and contour of sitting area (right)

3 Results

Publicly available datasets containing depth information focus on the detection of human activities, e.g. the UTKinect-Action Dataset [14], the Cornell Activity Dataset [12] or the DailyActivity3D [13]. The aim of these datasets is to detect human actions and activities and hence, different activities and actions are recorded in short sequences, while the person is standing in front of the sensor. On the other hand, scene understanding datasets (e.g. NYU Depth Dataset v2 [11], Berkeley 3-D Object Dataset [7]) does not contain tracking information since traditional scene understanding approaches are based on geometric information.

Our proposed approach neither focuses on the detection of actions within short sequences, nor on the incorporation of geometric information. In contrast, the introduced approach focuses on human behavior on the long-term and thus, no datasets are available. Hence, three different datasets (monitoring of a kitchen, living room and office environment) were recorded from a bird's eye view and is publicly available[2] in order to allow comparisons in the future. All scenes contain a sitting area as well as a walking area to be detected by the proposed approach.

[2] http://tracking-dataset.planinc.eu

The evaluation of the proposed approach 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.

Repeated sub-sampling validation is performed by randomly choosing the training set 6 times and averaging the results of all 6 runs. Figure 2 depicts the dependency of the f-score on the size of the training set (number of training days), where the sitting area of the kitchen is shown in the left picture and the walking area of the kitchen is shown in the right picture. The red dotted curve depicts the result of the unfiltered data, whereas the blue curve depicts the result if applying the filtering step as introduced in Section 2. The walking area is modeled with a resulting f-score of 0,87 for the filtered data and 0,84 for unfiltered data. However, the f-score for the sitting area achieved only 0,44 (unfiltered) respectively 0,33 (filtered).

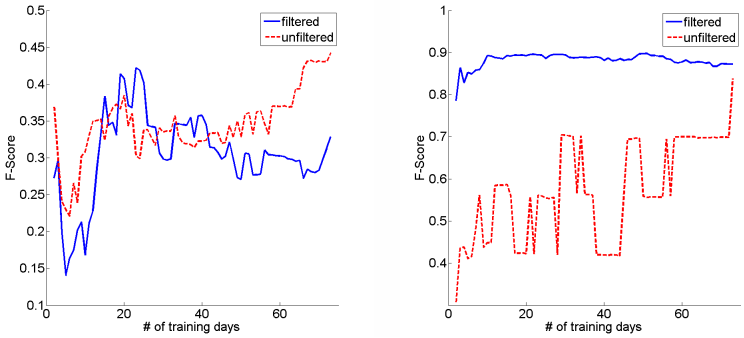

Fig. 2. F-score of the sitting (left) and walking area (right) in the kitchen dataset depending on the number of training samples (average of 6 runs)

In order to perform further analysis on these quantitative results, qualitative evaluation of the results is shown in Figure 3: during the 6 runs, two different models are obtained - the correct one is obtained four times (depicted on the left side) and a wrong one is obtained two times (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 much less training data, hence when using more training data, the influence of each training day is minimized. However, the low f-score is also a result of the ground truth labeling - as can be seen in the left image of Figure 3, the detected sitting area considers only one bench while the other bench is ignored since the second bench is already outside the tracking range. Moreover, the table is considered as sitting area in the ground truth and thus an f-score of 1 can not be achieved. Hence, ground truth annotation need to be adopted according to the range limits of the tracking algorithm and the table need to be excluded from the ground truth. However, qualitative analysis of the

walking area in the kitchen dataset shows that the walking area is modeled very well in both cases and thus result in an f-score of greater than 0,8. Moreover it is shown, especially in the walking area of the kitchen, that the filtering step is able to enhance the robustness of the proposed approach, since a more stable and higher f-score is achieved.

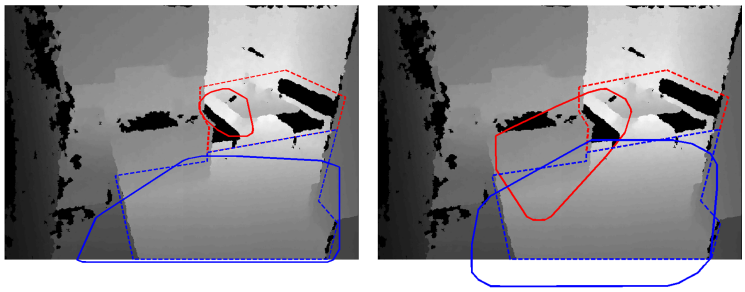

Fig. 3. Results of the scene understanding approach: sitting (red) and walking areas (blue) in the kitchen dataset with respective ground truth marked with dotted lines

The quantitative results of the living room and office dataset are depicted in Figure 4 and Figure 5: in contrast to the kitchen dataset, higher f-scores are achieved. The f-score for the sitting and walking area in the living room is 0,75 respectively 0,73, both for the filtered dataset. On the office dataset, f-scores of 0,68 for the sitting area and 0,88 for the walking area are achieved by using the maximum number of training data. The office dataset only consists 20 days of monitoring, hence the influence of increasing the training set can be seen since the f-score raises with adding additional training data. Moreover, the office dataset is not as challenging as the kitchen and living room training set since almost no tracking errors are present and thus, applying the filtering does not dramatically change the result. This is due the fact that the most common tracking errors experienced with the living room and kitchen dataset are the fitting of the skeleton to doors and other objects. In the office scene, no doors are within the field of view and thus enhancing the robustness of the tracker, yielding in better results already with little training data.

Qualitative analysis of the results show that the regions are modeled accurately and confirm the quantitative results: Figure 6 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.

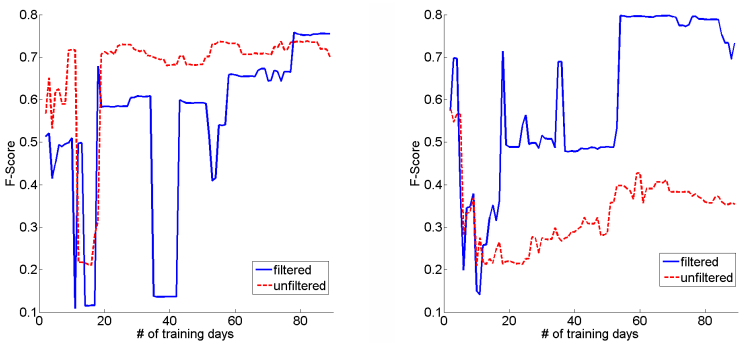

Fig. 4. F-score of the sitting (left) and walking area (right) in the living room dataset depending on the number of training samples (average of 6 runs)

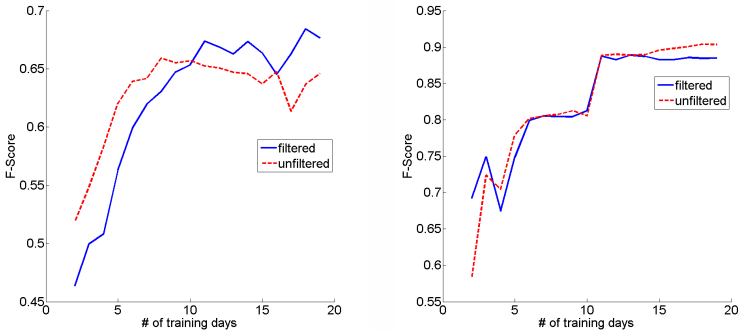

Fig. 5. F-score of the sitting (left) and walking area (right) in the office dataset depending on the number of training samples (average of 6 runs)

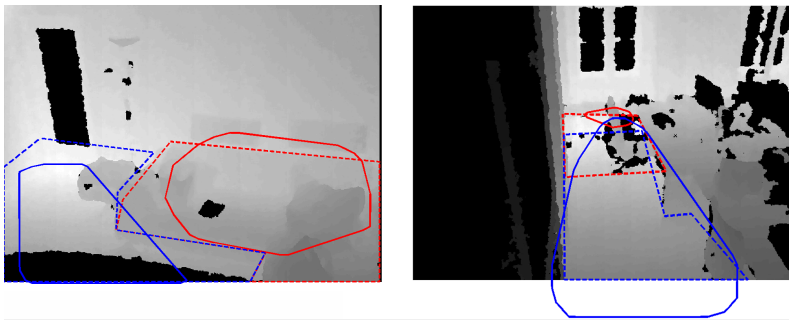

Fig. 6. Results of the scene understanding approach: living room (left) and office dataset (right) with respective ground truth marked with dotted lines

4 Conclusion

This paper introduced a novel method for human centered scene understanding, based on continuous tracking data obtained from a depth sensor. Quantitative and qualitative evaluation showed that the proposed approach is able to accurately model functional areas and thus showed that a human-centric approach is feasible for scene understanding. Due to possible occlusions, the limited field of view range of the sensor, only areas within the field of view can be covered by the system. However, in order to extend the proposed approach, more sensors can be used. Therefore, two different approaches to extend the range of the system are feasible: one generic behavioral model is obtained by fusing tracking data from different sensors and obtain one model for the whole scene. On the other hand, each sensor can obtain a separate behavior model for the corresponding field of view and only the results are combined.

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

References

1. OpenNI. http://www.openni.org (2011), [Online; accessed 10-April-2014]
2. Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A., Efros, A.: Scene semantics from long-term observation of people. In: ECCV. pp. 284–298 (2012)
3. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. PAMI 32(9), 1627–1645 (2010)
4. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People Watching: Human Actions as a Cue for Single-View Geometry. In: ECCV. pp. 732–745 (2012)
5. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3D scene geometry to human workspace. In: CVPR. pp. 1961–1968 (2011)
6. Gupta, S., Arbelaez, P., Malik, J.: Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In: CVPR. pp. 564–571 (2013)
7. Janoch, A., Karayev, S., Jia, Y., Barron, J., Fritz, M., Saenko, K., Darrell, T.: A category-level 3-d object dataset: Putting the kinect to work. In: ICCV Workshop on Consumer Depth Cameras for Computer Vision. pp. 141–165 (2013)
8. Lu, J., Wang, G.: Human-centric indoor environment modeling from depth videos. In: ECCV-Workshops and Demonstrations. pp. 42–51 (2012)
9. Mutch, J., Lowe, D.G.: Multiclass Object Recognition with Sparse, Localized Features. In: CVPR. vol. 1, pp. 11–18 (2006)
10. Planinc, R., Kampel, M.: Robust Fall Detection by Combining 3D Data and Fuzzy Logic. In: ACCV Workshop on Color Depth Fusion in Computer Vision. pp. 121–132. Daejeon, Korea (2012)
11. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760 (2012)
12. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from rgbd images. In: PAIR. pp. 842–849 (2011)
13. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. pp. 1290–1297 (2012)
14. Xia, L., Chen, C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: CVPR-Workshop. pp. 20–27 (2012)