Human Centered Scene Understanding based on Depth Information - How to Deal with Noisy Skeleton Data?

Rainer Planinc and Martin Kampel

Vienna University of Technology, Computer Vision Lab
Favoritenstrasse 9-11/183-2, A-1040 Vienna
{rainer.planinc, martin.kampel}@tuwien.ac.at

Abstract. Scene understanding is a challenging task and and mainly based on geometric or object centered approaches. Hence, the aim of this paper to introduce a novel human centered approach for scene analysis and tackle challenges of noisy long-term tracking data obtained by a depth sensor. Hence, fast filtering mechanisms are proposed to filter noisy tracking data, reducing the number of outliers and thus significantly improving the accuracy of the detection of walking and sitting areas within indoor environments. Evaluation is performed on two different scenes containing 18 and 34 days of tracking data and shows that detecting and filtering invalid tracking information dramatically increases the accuracy.

1 Introduction

In order to understand the structure of a scene, objects play an important role for humans since they offer different functions: beds are used by humans to lie on, chairs are used to sit on. Traditional scene understanding approaches are based on geometric information about the room and objects within the room [4, 7, 8, 10, 12]. However, object recognition is a big challenge in the area of computer vision [4]. But scene understanding can not only be seen as an object or geometric centered approach, but also as an human centered approach. Human centered approaches are not based on objects but on functionalities of objects, they are offering for humans [5]. Analyzing the person's actions (e.g. sitting, standing, walking) can be used to analyze the scene, the person is interacting with. Current user centered approaches combine information about the user together with geometric or object information.

This paper introduces the novel use of 3D long-term tracking based on depth information for obtaining object and scene functionalities within home environments, without incorporating geometric knowledge about the scene. Tracking is performed with a standard tracking algorithm provided by the OpenNI SDK [1]. This skeleton tracking algorithm is optimized for the detection and tracking of body parts and fitting a skeleton while the user is actively interacting with the Kinect sensor (i.e. the user is standing in front of the sensor). However, in the proposed approach of long-term tracking, the Kinect is placed in the corner of a room and used to track the movement throughout the room 24/7. Since the tracking algorithm is not optimized for this scenario, tracking errors occur and results in noisy and wrong tracking data. Hence, the contribution of this paper is to introduce filtering mechanisms in order to efficiently detect and eliminate wrong tracking data while preserving relevant information to define functional areas of humans, i.e. walking and sitting areas.

The rest of this paper is structured as follows: Section 2 summarizes related work. Challenges for 3D skeleton tracking algorithms within an home environment and methodologies for data filtering are introduced in Section 3. An evaluation of the proposed approach is presented in Section 4 and finally a conclusion is drawn in Section 5.

2 Related Work

Related work can be divided into two categories: object centered approaches and human centered approaches. Object and geometric centered approaches are traditionally used for scene understanding and thus the research within the field of object detection and classification is well established (e.g. [4, 7, 8, 10, 12]). However, the goal of this paper is the introduction of an human centered approach and thus, only related work within the field of human centered approaches are discussed.

Human centered approaches for scene understanding are introduced by Gupta et al. [6]. They introduce a proof of concept system focusing on the human workspace rather than on the object by combining single-view indoor geometry estimation and human pose analysis. The scene geometry is based on single images, where the occupied voxels are detected. Based on motion capture data, possible poses for a human (e.g. sitting) are extracted. In the last step, human scene interactions are modeled by analyzing the possible poses within the scene and thus recognizing free areas and surfaces supporting e.g. the sitting pose.

The work of Delaitre et al. [3] describes object functions within indoor scenes by the use of long-term tracking of humans, pose analysis and object appearance. The authors analyze the relation between human actions and objects. Pose analysis is restricted to the poses standing, reaching and sitting and is performed using the approach introduced by Yang and Ramanan [13]. Reliable pose estimation is still a problem in computer vision and contains noisy data [3]. The authors minimize the problem of noisy data by enhancing the number of people interacting with the object, i.e. analyzing a longer period of time using time-laps videos to receive reliable results. However, this method only works if most of the data does not contain noise and if enough training data is available. Together with appearance features, the person-object interaction withing an indoor scene is modeled. In contrast to Gupta et al. [6], Delaitre et al. [3] model areas where people are sitting, whereas Gupta et al. [6] model areas where people can sit theoretically.

Fouhey et al. [5] extends the approach of Delaitre et al. [3] by not only recognizing objects, but modeling the scene in 3D based on the object functionalities.

However, they also use both, appearance and human action information to obtain information about the scene. Analog to Delaitre et al. [3], Fouhey et al. [5] focus on the poses standing, sitting and reaching in order to classify surfaces into walkable, sitable and reachable surfaces. In the first step, poses are classified in time-lapse videos. Based on the poses, estimates of the functional surfaces (walking, sitting and reaching) are generated and combined with a 3D room geometry hypothesis to obtain a functional 3D scene description.

A proof of concept of human-centric scene modeling based on depth videos is introduced by Lu and Wang [9]. Since this paper is considered as proof of concept, analysis and evaluation of the algorithm is performed on data of only several minutes. Based on a depth video, the background is estimated and subtracted from each video frame to obtain the foreground (human silhouette). In combination with vanishing points estimation of the background image, a room hypotheses including supporting surfaces for human actions is created. The person is modeled as 3D cuboid and thus allows to estimate the free area of the room, where humans can walk. In combination with a pose estimation performed on the silhouette of the person, objects are modeled as 3D boxes within the scene. However, the authors of [9] 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 order to deal with noisy and defective skeleton data, Azimi [2] summarizes ways to smooth the skeleton tracking data by applying smoothing filters. Smoothing filters are applied to smooth the skeleton data over time, e.g. to remove jitter during the tracking. In contrast to other scenarios (e.g. rehabilitation), a highly accurate position of the skeleton joints is not required in the proposed approach, since small jitter of the joints does not influence the obtained scene functionalities. However, in the case of long-term tracking, tracking errors (i.e. an object is considered to be a person) are more important than jitter, but can not be filtered by applying smoothing filters since the skeleton itself is correct (i.e. does not contain jitter), but it is fitted to an object instead of a person.

3 Methodology

One of the aims of human centered scene understanding is the detection of areas within the room, offering different functions for humans. The proposed approach focus on the functions ”walking” and ”sitting”, since these are functions being able to describe the room efficiently: if an area is considered as walking area, no objects are present within this area (otherwise people could not walk there). If objects are present, people can either use them to relax by sitting or lying on them (e.g. bed, chair, sofa) or interact with them (e.g. table, wardrobe). Since the interaction with all possible types of objects is not in the scope of this paper, this work focus on the detection of ”walkable” and ”sitable” areas within a scene, being supported by the room layout. In contrast to related work, the proposed approach is purely based on 3D long-term tracking information from a depth sensor, hence no geometric information is used and no manual annotations (i.e. training) are needed in any step.

The workflow of the proposed approach is depicted in Figure 1: starting with 24/7 depth data, the skeleton joints are extracted by the use of OpenNI [1] and used as features for further processing. However, the proposed approach can be generalized in order to allow various tracker to be used since the processing pipeline does not change if another tracking algorithm is used. Moreover, the aim of this paper is not to develop new or improve existing tracking algorithms, but to deal with noisy and wrong tracking results. Based on the tracking data, the height of the human's center of mass (CoM) is calculated (with respect to the ground floor). Before walking and sitting areas are detected by the clustering of height data, the proposed filtering mechanisms are applied in order to remove outliers and eliminate tracking errors. Finally, the clustered tracking data indicate the areas where people are sitting and walking and thus allows to describe the scene.

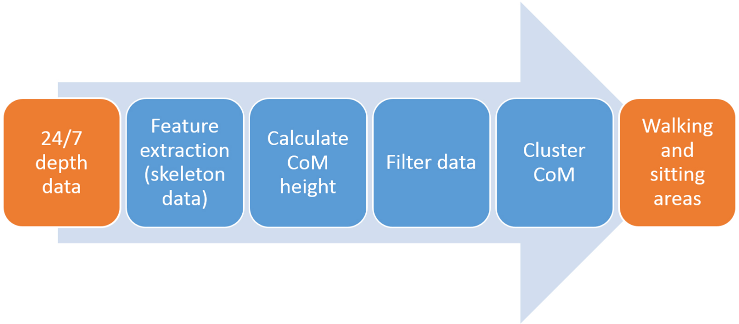


Fig. 1. Workflow of the proposed approach

The proposed approach is only based on indoor tracking information obtained by analyzing the skeleton of humans being tracked on the long-term range, i.e. over the duration of weeks and months. The main focus of this work is the introduction of fast and efficient filtering mechanisms in order to eliminate outliers being introduced due to the use of NITE [1], being responsible for the skeleton extraction and tracking. Since this tracker is optimized for the active interaction of the user with the sensor, as this is the case in commercial entertainment applications, tracking results contain jitter, are noisy or objects are tracked as humans by mistake (Figure 2). By the use of a filtering step as pre-processing and incorporating additional knowledge and constraints, the results of the obtained walking and sitting areas are dramatically improved. The filtering process is based on the following three criteria:

– **Tracking confidence** (provided by OpenNI): a confidence value is provided by OpenNI and is used to eliminate tracking errors - if tracking data is not robust (e.g. person is occluded), it is discarded. The influence of this value and its efficiency in removing outliers is analyzed in the evaluation section.

- **Body Orientation**: the orientation of the body indicates if the person is walking or sitting (based on the approach introduced by Planinc and Kampel [11]), since during walking and sitting an upright pose is assumed. The body orientation is calculated by the analysis of 3 body joints within the upper body with respect to the ground floor. A line is fitted to the center of the shoulder, the spine as well as the hip center and represents the orientation of the upper body. Hence, using this feature, all tracking information where the person is not in an upright pose is filtered out (including tracking errors where an object is wrongly tracked).
- **Person's height** (distance to the ground floor): when the CoM is tracked, the height of the CoM while a person is sitting or walking can be restricted, since it can be assumed that the distance is not more than 2 meters or less than 20 cm from the ground floor. This broad range of thresholds (valid data from 20 cm to 2 m) can be narrowed down by automatically learning the optimal thresholds - however, since these thresholds depend on the scene and application, the use of a wide range is proposed in order to not being too restrictive and provide a generalizable approach.

Although these criteria may be straightforward, applying the proposed filtering allows to eliminate most tracking errors, especially where parts of the furniture are tracked as humans and thus dramatically improves the results of the consecutive processing steps. An example of this tracking error is shown in Figure 2, where a kitchen door is tracked resulting in a skeleton being tracked on the kitchen top (although the person is far away from the ground floor).

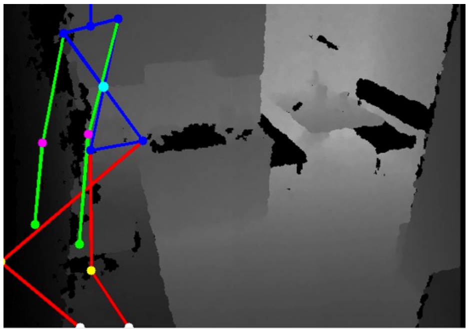


Fig. 2. Tracking error: furniture is tracked as a person

It should be noted that filtering outliers and focusing on only high qualitative data significantly reduces the amount of data since a high number of measurements is not considered. However, since the filtering is used for a long-term tracking approach, a high number of data points (over the duration of weeks or months) exist and thus, eliminating data points does not influence the result since still enough data points are available.

After filtering, the tracking data is clustered according to the height of the CoM (i.e. distance from the CoM to the ground floor) using the k-nearest neighbor algorithm to divide the tracking data (CoM) into a sitting and a walking class (k=2). Although tracking data does not only contain sitting or walking data, but also additional activities (e.g. picking something up from the floor, loading and unloading the dishwasher, etc.), it can be assumed that these activities are only minor activities (with respect to the duration) and that a walkable surface is mostly used for walking and that sofas and chairs are mainly used for sitting. Hence, when using long-term tracking information these minor activities can be ignored and only major activities (walking and sitting) remain. Hence, the use of long-term tracking data is feasible to indicate walking and sitting areas within an indoor environment, without geometric information or training data. Changing environments (e.g. moving chairs) are considered in two ways: first, temporarily changes (i.e. for a short period of time) does not influence the result, since long-term tracking data is obtained and thus tracking data represents the long-term behavior (i.e. is this are used as walkable area most of the time?). Second, permanent changes can be handled by an self adopting approach and using a rolling window and thus only considering the last x days/weeks, resulting in an adopted model.

4 Evaluation

Evaluation is based on the tracking data of one flat where two adults live in, where one sensor (Asus Xtion pro) is placed in the living room for 34 days, the second sensor in the kitchen for 18 days and tracking data is recorded using OpenNI [1]. Humans were present for around 3-4 hours per day during the week and approximately 6 hours per day during the weekend. Both scenes are depicted in Figure 3: in the living room, a sofa and free area is dominant where the free area is not only used to walk, but also during housecleaning. The kitchen scene depicts a kitchen including cupboards, a kitchen top and a cook-top. Moreover, benches and a table for meal intake are within the scene and thus the benches are regularly used. These specific field of views have been chosen in order to demonstrate the potential of the proposed approach - however, arbitrarily sensor positions are possible, as long as at least a free walking area is within the field of view. Since the approach is based on 3D data, the performance is independent from the location of the sensor.

In order to verify the accuracy, walking and sitting areas are annotated manually and the number of true positives, false positives, true negatives and false negatives are calculated to obtain the F1-score. Figure 4 depicts the kitchen area as example for the evaluation process: the sitting area is annotated in red, the walking area in blue. Clustered tracking data is visualized with yellow circles representing the tracking information belonging to the sitting cluster whereas

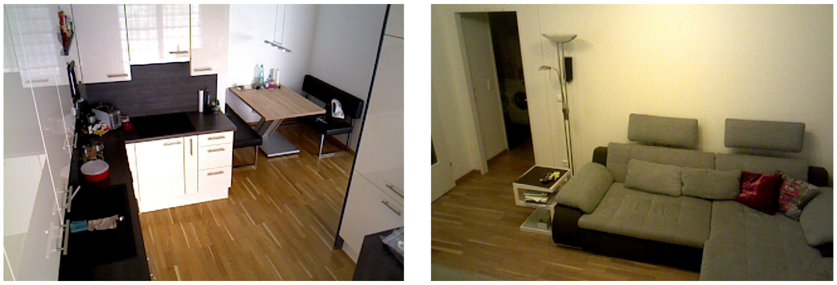


Fig. 3. Image of the analyzed scenes: kitchen (left) and living room (right)

cyan asterisks visualize tracking data of the walking cluster. Please note that the scenes are chosen since they are dynamic, i.e. objects are moved temporarily. Since the proposed approach is only human centered, these movements does not influence the result of the scene analysis.

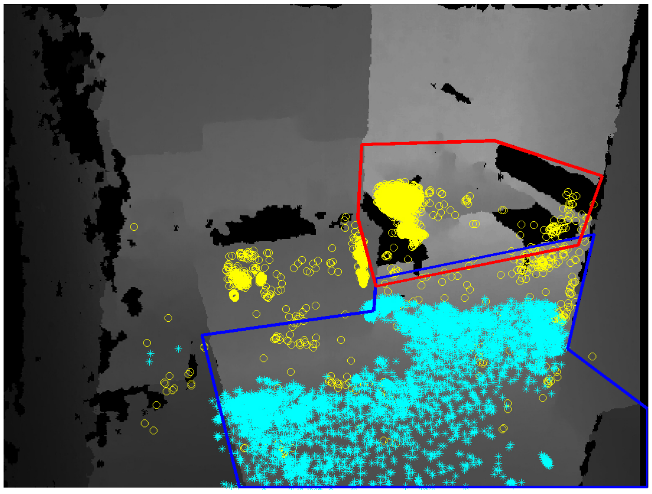


Fig. 4. Kitchen scene: ground truth annotation of sitting (red) and walking area (blue) and center of mass tracking points clustered to walking (cyan asterisk) and sitting (yellow circle)

The evaluation is performed on the kitchen and the living room separately, since different parameter settings have different implications, depending on the scene. Figure 5 illustrates the influence of the body orientation, the use of a height threshold as well as the use of "confident" (marked by OpenNI) data only. The value of body orientation is calculated according to [11], where the

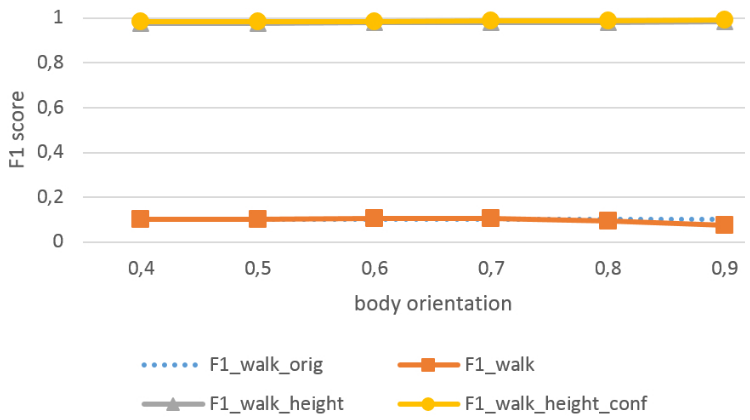


Fig. 5. F1 score for the evaluation of the walking area in the kitchen depending on the threshold for the body orientation

similarity of the orientation to the ground plane is calculated based on the angle between ground plane and upper body. The range of the body orientation index is from 0 (parallel to the ground floor) to 1 (orthogonal to th ground floor).

As can be seen clearly, only using upright positions (body orientation) does not influence the result of the kitchen scene, since the F1 score is constant over time. However, the use of a height threshold (F1 walk height) dramatically increases the performance of the system from an F1 score of 0.1 to an F1 score of 0.98. This dramatically increase of performance is gained due to eliminating many wrong tracks as depicted in Figure 2. These wrong tracks occurred due to the movement of the cupboard doors, which are wrongly tracked by OpenNI resulting in tracks on the kitchen top with a high distance to the ground floor. Hence, these wrong tracking results can be filtered out using the proposed approach easily, since the height of the CoM is much higher than 2 meters.

On the other hand, the dependence of the body orientation threshold in the living room scene is depicted in Figure 6: a clear influence of the body orientation threshold is seen since the F1 score is higher, when using only strict "upright" poses having a body orientation index of 0.8 or higher. Moreover, the use of only confident joint values (marked by OpenNI) is also recommended, since this combination leads to the best results. Please note that only figures for the class "walking" are shown in this paper since the results for the class "sitting" are analog and thus does not need to be discussed further, but are shown in the summary in Table 1.

Table 1 summarizes the results and the influence of the evaluated parameters. A range of F1 score in this table indicates that the results are within the specified range and depend on the defined threshold. As can be clearly seen, the use of thresholds for eliminating outliers significantly increases the accuracy of the detection of walking and sitting areas. Due to the use of fast filtering mechanisms,

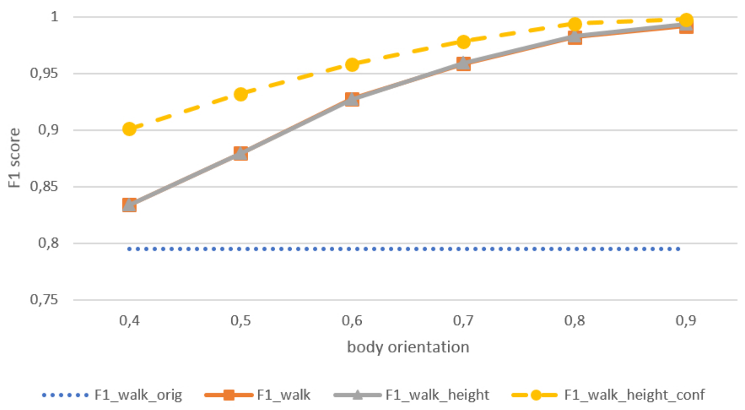
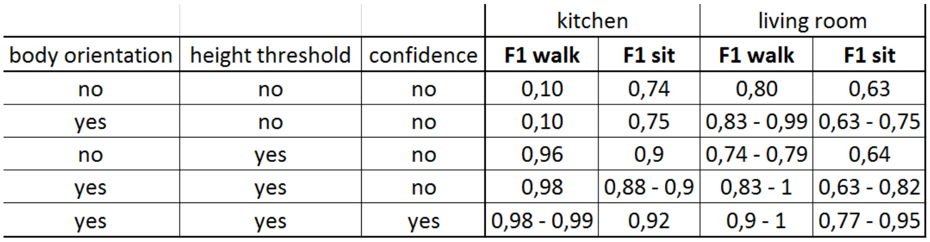


Fig. 6. F1 score for the evaluation of the walking area in the living room depending on the threshold for the body orientation

Table 1. F1 score indicating the influence of different parameter combinations



			kitchen		living room	
body orientation	height threshold	confidence	**F1 walk**	**F1 sit**	**F1 walk**	**F1 sit**
no	no	no	0,10	0,74	0,80	0,63
yes	no	no	0,10	0,75	0,83 - 0,99	0,63 - 0,75
no	yes	no	0,96	0,9	0,74 - 0,79	0,64
yes	yes	no	0,98	0,88 - 0,9	0,83 - 1	0,63 - 0,82
yes	yes	yes	0,98 - 0,99	0,92	0,9 - 1	0,77 - 0,95

the performance of the classification can be improved, without increasing the computational demands since only thresholds are applied.

5 Conclusion

This paper introduced a human centered approach for scene understanding based on long-term noisy skeleton by using a depth sensor. To achieve reasonable results, skeleton data need to be pre-processed and filtered since many tracking errors are present. Due to the incorporation of prior knowledge, the accuracy of the tracking data is enhanced dramatically. This results in an enhanced accuracy of the detected walking and sitting areas. The proposed approach is generalizable and can be applied to different indoor areas, where tracking results (e.g. obtained by OpenNI) are noisy and contain wrong tracking information (i.e. objects are wrongly tracked). The proposed filtering mechanisms are able to filter tracking data without the need for explicit training - however, this approach focus on long-term tracking data, hence, the more data available, the more accurate the results. However, it was shown that only tracking information from a few days already yields in reasonable results for the detection of walking and sitting areas within flats. Moreover, the proposed filtering approaches can also be used as a pre-processing step before applying other methods, since the filtering step corrects the data and removes outlier from the tracking data. Future work deals with the further development of the human centered approach in scene analysis in order to introduce a flexible and robust approach for scene understanding, without considering geometric constraints.

References

1. OpenNI. http://www.openni.org, 2011. [Online; accessed 10-April-2014].
2. M. Azimi. Skeletal Joint Smoothing. http://msdn.microsoft.com/en-us/library/jj131429.aspx, 2012. [Online; accessed 10-April-2014].
3. V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 284–298, Florence, 2012.
4. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, Sept. 2010.
5. 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 *Proc. of the European Conference on Computer Vision (ECCV)*, pages 732–745. Springer Berlin Heidelberg, 2012.
6. A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3D scene geometry to human workspace. In *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1968. IEEE, June 2011.
7. S. Gupta, P. Arbelaez, and J. Malik. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 564–571, 2013.
8. D. Holz, S. Holzer, R. B. Rusu, and S. Behnke. Real-Time Plane Segmentation Using RGB-D Cameras. In *Proc. of the RoboCup: Robot Soccer World Cup*, number D, pages 306–317. Springer Berlin Heidelberg, 2012.
9. J. Lu and G. Wang. Human-centric indoor environment modeling from depth videos. In *Proc. of European Conference on Computer Vision (ECCV) - Workshops and Demonstrations*, pages 42–51, 2012.
10. J. Mutch and D. G. Lowe. Multiclass Object Recognition with Sparse, Localized Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 11–18, 2006.
11. R. Planinc and M. Kampel. Robust Fall Detection by Combining 3D Data and Fuzzy Logic. In *ACCV Workshop on Color Depth Fusion in Computer Vision*, pages 121–132, Daejeon, Korea, 2012. Springer.
12. G. Tsai and B. Kuipers. Real-time indoor scene understanding using Bayesian filtering with motion cues. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 121–128. IEEE, Nov. 2011.
13. Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011.