Feasibility of Human Activity Recognition Using Wearable Depth Cameras

Philipp Voigt, Matthias Budde, Erik Pescara
TECO, KIT, Germany
{pvoigt,budde,pescara}@teco.edu

Manato Fujimoto, Keiichi Yasumoto
Ubi-lab, NAIST, Japan
{manato,yasumoto}@is.naist.jp

Michael Beigl
TECO, KIT, Germany
michael@teco.edu

ABSTRACT

Human Activity Recognition (HAR) with body-worn sensors has been studied intensively in the past decade. Existing approaches typically rely on data from inertial sensors. This paper explores the potential of using point cloud data gathered from wearable depth cameras for on-body activity recognition. We discuss effects of different granularity in the depth information and compare their performance to inertial sensor based HAR. We evaluated our approach with a total of sixteen participants performing nine distinct activity classes in three home environments. 10-fold cross-validation results of KNN and Random Forests classification exhibit a significant increase in F-score from inertial data to depth information (by > 12 percentage points) and show a further improvement when combining low-resolution depth matrices and sensor data. We discuss the performance of the different sensor types for different contexts and show that overall, depth sensors prove to be suitable for HAR.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; I.5.m. Pattern Recognition: Miscellaneous

Author Keywords

Human Activity Recognition; Mobile Computing; Google Project Tango; Wearables; Depth Sensor.

INTRODUCTION & RELATED WORK

Human activity recognition (HAR) in home environments has been widely explored in recent years. Prominent applications include health care and assisted living for the elderly, home automation and energy saving [8, 9]. Inertial sensors in smartphones or attached to different parts of the body have been widely used for HAR [2]. Filippoupolitis et al., for instance, achieved promising results relying only on accelerometer data from smartwatches [7]. Other related work considers, i.e., the varying location of a smartphone on the body [5] or postural transitions [15]. They struggle, however, with activities of similar postures and movement patterns and good results are often achieved only for a small number of high-level activities. This work addresses these limitations by proposing HAR using body-worn depth cameras.

Initially, activity recognition with cameras has been performed with classic computer vision approaches involving image processing, object detection, and recognition algorithms [11]. More recently, low-cost structured light depth sensors such as the Microsoft Kinect enabled research based on 3D point cloud data. These approaches commonly use 3rd-person video, limited to a fixed viewpoint [4] and location. Other projects used depth cameras mounted on mobile robots following the user [10]. Egocentric video captured from cameras mounted directly on the user's head or body observes the scene from a 1st-person viewpoint. Cho et al. embedded an image sensor and an accelerometer into a belt buckle to detect the direction of movement, such as *Walking Forward* or *Turning* based in optical flow and acceleration data [6]. Pirsiavash and Ramanan recognize detailed Activities of Daily Living using a chest-mounted video camera [12]. With similar hardware, Castro et al. were able to recognize 19 specific activities using a Convolutional Neural Network [3]. Rather than image data, we propose the use of depth information which – to the best of our knowledge – is a novel approach that does not rely on computationally intensive object recognition and is suitable for real-time application.

SYSTEM DESIGN

For our study, we employ the Google Project Tango platform (*Lenovo Phab 2 Pro*, see Figure 1) which offers both a depth sensor and inertial sensors in the portable form factor of a smartphone. It has gained popularity in recent years [1] and has been shown to exhibit an average error similar to popular desktop sensors [13].

Figure 1. We used a body-worn Tango system to investigate Human Activity Recognition (HAR) using the depth information from the camera.

Recording and Labeling

The built-in wide-angle and time-of-flight depth cameras together with its inertial measurement unit enable the device to calculate relevant features about the environment (area learning), localize itself within it (motion tracking) and generate 3D point clouds of it (depth perception). The device allows flexible programming and is portable and small enough to be carried on-body. These factors made it a convenient platform for this study. We developed an Android application that simultaneously records point cloud data and inertial sensor data on the same device. Ground truth labels can be assigned within the application before recording. To minimize distraction while collecting data, we developed a remote control application running on a second smartphone which can be used to assign labels as well as start and stop recordings.

Segmentation

The recorded data was then preprocessed by applying segmentation with a segment size of 1 second on both sensor and depth data. This segment size was found to be useful for sensor data [7]. Furthermore, the Tango API provides 1-3 point clouds per second, depending on the number of points identified in the scene. We only consider the first point cloud in each segment, since merging or averaging all point clouds per segment only increased processing time without improving the results. Similarly, expensive filtering and outlier removal algorithms did not impact the performance and were not applied in this study.

Temporal and Visual Feature Extraction

For each of the segments, we extracted mean and standard deviation as basic temporal features for all base and composite sensors listed in Table 1.

Our first set of visual features comprises four *Depth Histograms* with equal bin width (*eql*), as well as square (*sqr*), cubic (*cub*) and exponentially (*exp*) scaled bins. The nonlinear binning strategies aim to emphasize the area closer to the device. The reason is that activities are often most clearly distinguishable by the objects that the user is holding or interacting with. Bins are defined by break points that separate bins from one another. Let n be the number of bins, $i \in [1,2,...,n]$ the break point indices, d the maximum distance and a the exponential growth factor, we calculate the break points according to the functions $eql_i = \frac{i \times d}{n}$, $sqr_i = \frac{i^2 \times d}{n^2}$, $cub_i = \frac{i^3 \times d}{n^3}$ and $exp_i = \begin{cases} 0, & \text{if } i = 0 \\ \frac{a^i \times d}{a^n}, & \text{otherwise.} \end{cases}$

Secondly, we generate a *Depth Matrix* by splitting each point cloud into a 3×3 and 4×4 grid of equal sized cells. First we calculate the angles that the viewing direction forms with the directional vector from camera origin to each point: $\alpha = atan(\frac{z_i}{x_i}) - atan(1)$ and $\beta = atan(\frac{z_i}{y_i}) - atan(1)$. Based on these two angles, we can assign each point to a row and column of the matrix. Finally, we use the mean distance of points in each cell as features (see Figure 2). We used this approach since similar distance-based features for the entire point cloud did not provide enough distinction between activities.

Sensor	Type	Dimensions
Accelerometer	Base	3
Gyroscope	Base	3
Linear Acceleration	Composite	3
Gravity	Composite	3

Table 1. Sensor data recorded by our application. The Android API provides data taken from individual physical sensors (Base Sensors) as well as values derived from multiple sensors (Composite Sensors).

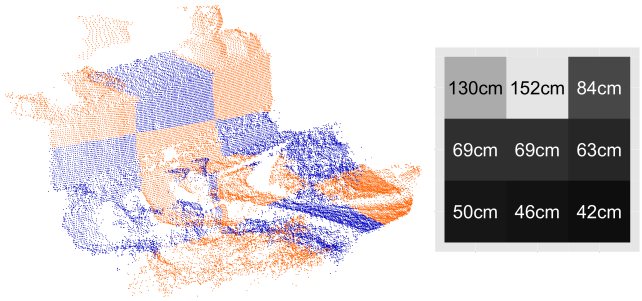

Figure 2. Visualization of a 3×3 grid applied to a sample point cloud of the kitchen counter (left) with the corresponding depth matrix (right). Each cell of the depth matrix contains the mean distance of all points in the corresponding orange or blue segment.

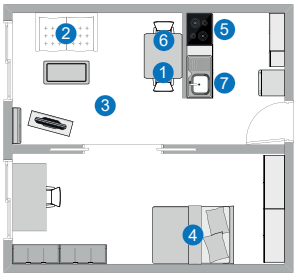

Figure 3. Floor plan of the smart home facility. Numbers 1–7 indicate the location of each activity listed in Table 2.

Sliding Window

The visual features above characterize one individual point cloud. To capture dynamic changes over time, we applied a 10-second sliding window in increments of 1 sample. To keep the approach suitable for real-time application, we only consider past data. As a standard measure for the similarity of histograms, we used chi-square distance between consecutive samples and calculated mean and standard deviation for the 10-second window. For depth matrices, we derived the same measures directly from the mean distances of each cell.

EVALUATION

To evaluate our approach, we collected a set of labeled training data in a smart home facility operated by the Ubiquitous Computing Systems Laboratory at Nara Institute of Science and Technology [16] and in two private homes (see Figure 3). Each activity was carried out three times for one minute by 16 participants (1 female), aged 23 to 32 years ($\mu = 27.5$, $\sigma = 3.4$), with an average body height of 1.78 m ($\sigma = 0.1$). This provided us with a balanced dataset of 7.2 hours. We chose nine activities, as listed in Table 2 with their respective location in the smart home. The Tango device was strapped to the participant's chest keeping location and angle consistent without restricting natural movement of the arms. It was

	Label	Activity	Location
1	PC	Working on a laptop	1
2	TV	Watching TV on the couch	2
3	Book	Reading a book or magazine	2
4	Phone	Operating a smartphone	2
5	Cleaning	Vacuuming the living room	3
6	Sleeping	Lying in bed on the back	4
7	Cooking	Preparing eggs in a pan	5
8	Eating	Eating the eggs with a fork	6
9	Dishes	Washing the dishes	7

Table 2. Overview of recorded activities and their respective location in the apartment (see floor plan in Figure 3).

Features	Leave-Recording-Out	Leave-Subject-Out	Leave-Environment-Out
Inertial Sensors (38 features)	78.2%	50.8%	47.4%
Cubic Histogram (64 bins)	**94.6%**	**74.8%**	55.2%
VFH Histogram (31 bins)	87.6%	71.4%	51.2%
ESF Histogram (40 bins)	78.4%	61.3%	28.5%
Depth Matrix (4×4)	93.1%	74.0%	**60.6%**
Sensors & Histogram	**96.5%**	76.4%	63.8%
Sensors & Matrix	93.9%	**78.0%**	**69.4%**
All (Sensors, Histograms & Matrix)	95.9%	75.8%	66.8%

Table 3. F-Score of Random Forest classifiers: cross-validation using dynamic features calculated over a 10s sliding window. The results of the K-Nearest Neighbors (KNN) approach showed similar relative improvements and slightly lower absolute scores. Among the four histogram binning strategies, the cubic histogram was the most successful.

mounted in portrait orientation to maximize the vertical field of view since capturing both the user's interaction with objects in the lower part of the frame and the structure of the background in the upper region provided significantly better results. The data was stored with preassigned labels and recordings were started and stopped remotely.

Results

We trained a series of K-Nearest Neighbors (KNN) and Random Forest (RF) classifiers on the feature sets described above. As RF shows overall better performance than KNN, only RF values are presented here. Distance-based classification is common for histogram data, and we expected a tree-based classifier to separate activities based on specific depth intervals especially in depth matrices. We evaluated the classifiers with 10-fold cross-validation on participants (Leave-Subject-Out) and recordings (Leave-Recording-Out), and with 3-fold cross-validation for the three sites (Leave-Environment-Out). Table 3 shows the results of RF classification. Depth histograms and depth matrices provide discriminative features for the tested 9 activities. Across all three cases, the use of depth information increased the F-score by at least 12 p.p. (percentage points) compared to inertial sensors. The cross-validation also shows lower variance when using depth data.

Finally we compared these results to Viewpoint Feature Histograms (VFH) and Ensemble of Shape Functions (ESF), two histogram-based global point cloud descriptors implemented in the Point Cloud Library [14]. The two descriptors were chosen as they are fast enough for real-time application. However, both showed a performance lower than our proposed simple features with an F-score of up to 87.6% for VFH.

	Cleaning	Cooking	Dishes	Eating	Book	Sleeping	Phone	TV	PC
Cleaning	69.5	5.2	12.2	2.3	0.0	0.0	0.0	0.0	0.3
Cooking	12.5	65.2	13.8	19.6	0.8	<0.1	2.3	<0.1	7.8
Dishes	16.5	11.4	70.3	6.0	0.2	0.0	0.2	<0.1	0.8
Eating	1.3	8.0	3.1	53.5	6.3	0.0	2.9	0.2	11.9
Book	<0.1	1.6	0.1	3.3	71.5	0.0	18.0	2.9	1.2
Sleeping	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Phone	<0.1	2.3	<0.1	2.6	17.0	0.0	68.4	1.9	7.9
TV	0.0	<0.1	0.0	0.0	3.3	0.0	2.1	94.4	2.0
PC	0.1	6.3	0.4	12.7	0.9	0.0	6.1	0.6	68.0

	Cleaning	Cooking	Dishes	Eating	Book	Sleeping	Phone	TV	PC
Cleaning	96.5	0.1	0.0	0.0	0.3	<0.1	0.0	1.4	0.0
Cooking	<0.1	95.3	3.5	1.1	0.4	0.0	0.2	<0.1	1.2
Dishes	0.0	4.1	95.2	0.3	0.9	0.0	0.3	0.0	0.5
Eating	0.0	0.3	0.7	91.8	2.8	0.0	0.8	0.0	6.0
Book	0.0	<0.1	0.3	2.1	82.3	0.0	4.9	0.0	13.4
Sleeping	<0.1	0.0	0.0	0.0	0.0	100	0.0	<0.1	0.0
Phone	0.0	<0.1	0.0	1.1	3.5	0.0	91.6	0.0	2.4
TV	3.4	<0.1	0.0	0.0	0.0	0.0	0.0	98.5	0.0
PC	0.0	<0.1	0.3	3.6	9.7	0.0	2.2	0.0	76.4

Figure 4. Confusion matrices (Random Forests, Leave-Recording-Out) for inertial sensors (top) respectively for 4×4 depth matrix (bottom). Highest inertial sensor misclassification occurred between *Book* and *Phone*, and *Cleaning*, *Cooking* and *Dishes* which represent similar posture and movement. When using the depth matrix, classification error was significantly reduced.

Leave-Recording-Out

When the subjects are known, the system exhibits an F-score of up to 94% using only depth data. The results of four 64-bin histograms, on the one hand, show that emphasizing the area near the body to a certain degree is beneficial with the cubic strategy showing the best results. The equal-width histogram (F-score 91.2%) neglects small differences near the body while the exponential histogram (F-score 87.8%) lacks detail in greater distance, both leading to a lower F-score. Depth matrices, on the other hand, exhibit similarly high F-scores of >90% relying solely on nine or sixteen mean depth features. Dynamic features improved the performance slightly. However, using static features alone would allow the sensor to operate at a much lower sampling rate compared to inertial sensors. The confusion matrices in Figure 4 show that depth data is superior particularly in separating activities with similar posture (*Phone* and *Book*) and similar movement (*Cleaning*, *Cooking* and *Dishes*).

Leave-Subject-Out

When testing with unknown subjects, depth data shows a significant advantage over inertial sensors. A 4×4 depth matrix outperforms interial sensor data based classifiers by 23p.p. with an F-score of 74% (see Table 3). We can argue, that posture and body movement are unique to each subject. Thus information about the test subject is required when training a classifier on inertial sensor data. The environment, however, remains the same for all subjects, allowing our approach to succeed.

Leave-Environment-Out

When environment and subjects are unknown to the system, absolute F-scores are lower, but depth data still scores 13 p.p. higher than inertial sensors. While the smart home facility was a western-style home, other apartments were equipped with Japanese-style furniture. Therefore, the activities *TV*, *Reading*, *Eating*, *Phone* and *Book* took place sitting on the

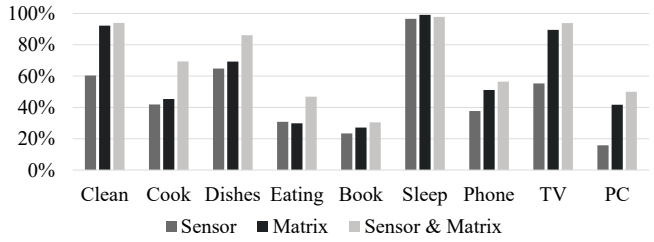

Figure 5. Per-class accuracy of leave-environment-out cross-validation.

Tatami floor with a very different point-of-view and posture. The results show that our approach is more robust against significant differences in the environment. Figure 5 compares the per-class F-scores. A 4×4 depth matrix reduced misclassification most notably among the stationary activities *PC*, *Phone* and *TV* and for the dynamic activity *Cleaning*. The latter was frequently confused with the other dynamic activities *Cooking* and *Dishes*. The major advantage of our approach is its ability to distinguish activities based on size and distance of objects the user is holding rather than posture and movement alone. Finally, a combination of depth matrices and inertial sensor data further improves the scores of dynamic activities by taking subtle differences in movement patterns into account.

Discussion

Our results clearly show the potential of wearable depth cameras for HAR. The Tango device allowed us to gain these initial insights without needing specialized hardware. Due to its size and weight, however, it is not comfortable enough to be worn over long periods of time. Furthermore, the device was designed for performance rather than energy efficiency which makes it unsuitable for continuous operation. Ultimately, we envision a compact, light-weight and unobtrusive depth sensor with a low resolution (e.g., 4×4 pixels) and frame rate (e.g., 1 min). We believe such a sensor in the shape of a brooch or shirt button could perform at a similar accuracy as the presented results while consuming significantly less energy.

CONCLUSION AND FUTURE WORK

In this paper, we have shown that a wearable depth camera can be used for human activity recognition. Our proposed approach requires only one single depth sensor which could operate at low sampling rates. With depth matrices, we were able to achieve >90% accuracy with a total of only nine features, which means that a low-resolution sensor in a compact form-factor would be sufficient. In a direct comparison, our approach outperformed inertial sensors and established point cloud descriptors for activities with similar posture and movement and showed lower variance overall. The simplicity of the proposed features makes our approach suitable for real-time and on-device feature extraction. Furthermore, our approach exhibited satisfying results even for unknown subjects and diverse environments showing a clear advantage over inertial sensors. Future work will include a long-term study in uncontrolled environments across several homes. We will also address the use of alternative depth sensors, their placement on the body, and their energy efficiency.

ACKNOWLEDGMENTS

Supported by the Baden-Württemberg-STIPENDIUM and JSPS KAKENHI grants JP16H01721 & JP17KT0080.

REFERENCES

1. M. Abdelaal, D. Reichelt, F. Dürr, K. Rothermel, L. Runceanu, S. Becker, and D. Fritsch. 2018. ComNSense: Grammar-Driven Crowd-Sourcing of Point Clouds for Automatic Indoor Mapping. *IMWUT* 2, 1 (2018).
2. A. Bulling, U. Blanke, and B. Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Comput. Surv.* 46, 3 (2014).
3. D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. 2015. Predicting Daily Activities from Egocentric Images Using Deep Learning. In *ISWC'15*. ACM.
4. C. Chen, R. Jafari, and N. Kehtarnavaz. 2016. A Real-Time Human Action Recognition System Using Depth and Inertial Sensor Fusion. *Sensors* 16, 3 (2016).
5. Y. Chen and C. Shen. 2017. Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition. *IEEE Access* 5 (2017).
6. Yongwon Cho, Yunyoung Nam, Yoo-Joo Choi, and We-Duke Cho. 2008. SmartBuckle: Human Activity Recognition Using a 3-axis Accelerometer and a Wearable Camera. In *HealthNet'08 Workshop*.
7. A. Filippoupolitis, B. Takand, and G. Loukas. 2016. Activity Recognition in a Home Setting Using Off the Shelf Smart Watch Technology. In *IUCC-CSS'16*.
8. W. S. Lima, E. Souto, T. Rocha, R. W. Pazzi, and F. Pramudianto. 2015. User activity recognition for energy saving in smart home environment. In *ISCC '15*.
9. R. Lutze and K. Waldhör. 2017. Personal Health Assistance for Elderly People via Smartwatch Based Motion Analysis. In *ICHI'17*.
10. K. Nakahara, H. Yamaguchi, and T. Higashino. 2016. In-home Activity and Micro-motion Logging Using Mobile Robot with Kinect. In *Mobiquitous'16*.
11. K. Ohnishi, A. Kanehira, A. Kanezaki, and T. Harada. 2015. Recognizing Activities of Daily Living with a Wrist-mounted Camera. *CoRR* (2015).
12. H. Pirsiavash and D. Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *CVPR'12*.
13. R. Roberto, J. P. Lima, T. Araújo, and V. Teichrieb. 2016. Evaluation of Motion Tracking and Depth Sensing Accuracy of the Tango Tablet. In *ISMAR'16*.
14. R. B. Rusu and S. Cousins. 2011. 3D is here: Point Cloud Library (PCL). In *ICRA'11*.
15. M. T. Uddin, M. M. Billah, and M. F. Hossain. 2016. Random forests based recognition of human activities and postural transitions on smartphone. In *ICIEV'16*.
16. K. Ueda, M. Tamai, and K. Yasumoto. 2015. A method for recognizing living activities in homes using positioning sensor and power meters. In *PerCom'15*.