CommonSense: Collaborative learning of scene semantics by robots and humans

Stefano Rosa
University of Oxford
Oxford, UK
stefano.rosa@cs.ox.ac.uk

Andrea Patanè
University of Oxford
Oxford, UK
andrea.patane@cs.ox.ac.uk

Xiaoxuan Lu
University of Oxford
Oxford, UK
xiaoxuan.lu@cs.ox.ac.uk

Niki Trigoni
University of Oxford
Oxford, UK
niki.trigoni@cs.ox.ac.uk

ABSTRACT

The recent introduction of robots to everyday scenarios has revealed new opportunities for collaboration and social interaction between robots and people. However, high level interaction will require semantic understanding of the environment. In this paper, we advocate that co-existence of assistive robots and humans can be leveraged to enhance the semantic understanding of the shared environment, and improve situation awareness. We propose a probabilistic framework that combines human activity sensor data generated by smart wearables with low level localisation data generated by robots. Based on this low level information and leveraging colocation events between a user and a robot, it can reason about semantic information and track humans and robots across different rooms. The proposed system relies on two-way sharing of information between the robot and the user. In the first phase, user activities indicative of room utility are inferred from consumer wearable devices and shared with the robot, enabling it to gradually build a semantic map of the environment. This will enable natural language interaction and high-level tasks for both assistive and co-working robots. In a second phase, via colocation events, the robot is able to share semantic information with the user, by labelling raw user data with semantic information about room type. Over time, the labelled data is used for training an Hidden Markov Model for room-level localisation, effectively making the user independent from the robot.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

Human-Robot Interaction, Wearable Devices, Localisation

IoPARTS'18, June 10, 2018, Munich, Germany

ACM ISBN 978-1-4503-5843-9/18/06…$15.00
https://doi.org/10.1145/3215525.3215526

ACM Reference Format:
Stefano Rosa, Andrea Patanè, Xiaoxuan Lu, and Niki Trigoni. 2018. CommonSense: Collaborative learning of scene semantics by robots and humans. In *IoPARTS'18: 1st International Workshop on Internet of People, Assistive Robots and ThingS , June 10, 2018, Munich, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3215525.3215526

1 INTRODUCTION

High-level semantic understanding of the environment is still an open problem for complex cyber physical systems involving robots and people. We envision that in the next five years, such systems will become ubiquitous. We consider applications where assistive robots are operating in domestic environments up to now almost exclusively inhabited by humans. In such environments, concepts such as room types and activities are important, not only because of the interaction with humans but also for abstracting spatial knowledge.

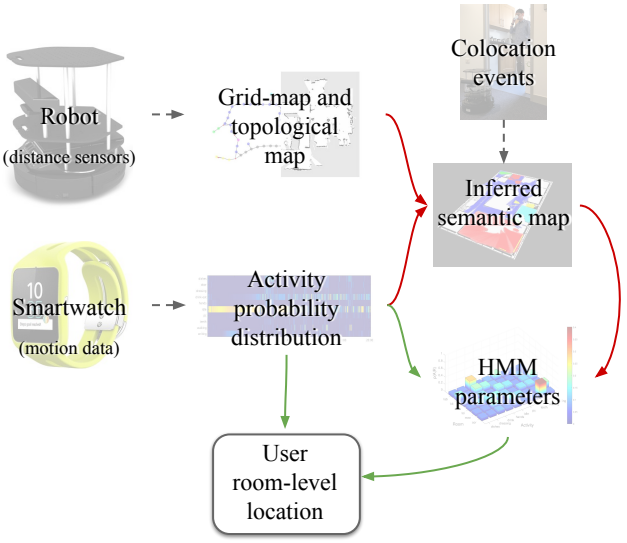


Figure 1: Architecture of the proposed system. Red arrows represent the semantic map inference information flow; green arrows represent the user localisation flow.

Robots typically perceive space in terms of gridmaps, topological or feature maps. Recent work has motivated the need for a high level understanding of the environment (e.g., semantic, affordances or high-level geometry) in order to enable emerging robotics applications [3]. Vision-based techniques for semantic mapping are well studied. [10] presents a conceptual model for semantic map representation, with different levels of abstraction, from sensor data to concepts, such as rooms, with associated appearance, and detected objects.

Semantic information has been integrated with dense 3D SLAM techniques such as KinectFusion in order to obtain a 3D semantic map of the environment. [13] proposes a semantic mapping approach for inferring room types using visual place categorisation.

These techniques tend to be very sensitive to the environment and require careful training and/or fine-tuning for each environment. We propose that semantic information should be spontaneously and effortlessly acquired by robots as a result of them interacting with humans.

On the other hand, locating people indoor using wearable devices, such as smartphones or smartwatches, requires either bespoke sensor infrastructure (e.g. WiFi, UWB [2], Bluetooth Low Energy (BLE) [15]) or, in the case of infrastructure-less methods, extensive offline training. Moreover, knowing the precise location of the user is not always useful for high-level interaction. Wrist mounted devices can also extract high-level information in the form of human activities.

Human activity recognition with inertial sensors has been well studied [12] [11] [14]. Activity classification has also been used as part of Simultaneous Localisation and Mapping frameworks. In [8] the authors proposed a 3D SLAM algorithm for users wearing wearable sensors, by including detected activities as landmarks in a particle filter SLAM approach. In [7] the approach is combined with semantic SLAM with the goal of adding robustness to errors in activity recognition. However, in both approaches the user carries several inertial sensors (wrist-mounted, hip-mounted, foot-mounted IMUs), and the coexistence with robots is not explored. Previous work on combining user activities with WiFi and acoustic data to localise users at room-level in domestic environments required a labour-intensive fingerprinting procedure to develop the WiFi map [9].

Differently from above work, we move away from location-based training efforts, and rely on lifelong learning from human-robot interactions. The idea is to progressively make wearable devices aware of their environment, thru the side-channel information provided by robots. To this end, we propose a system that enables robots and wearable devices to have a semantic understanding of their environment via colocation and interaction with each other. We believe that this is key to a variety of applications, from issuing simple commands to robots such as "Go to the kitchen", to tasks of collaborative nature like "The robot should go to the kitchen when the user (her smart wearable) is there".

The first intuition behind our approach is that user activities provide informative hints about the utility of each room. For example, a bedroom can be easily identified if people often sleep in that room. Once we address the problem of semantic mapping, it paves the way for inferring the sequence of room types that human devices traverse. A robot, who is now aware of semantic room labels, can *teach* human mobile devices how to recognise them based on their own signals. Specifically, we show how a robot can help mobile devices to tune the parameters of the graphical model that they use for localisation.

Semantic mapping and semantic localisation are two faces of the same coin; we address both by leveraging opportunistic *colocation* events between robots and human-held devices. Through the diverse lenses of robots and wearable devices, we show that they can both develop a semantic understanding of their space.

We finally validate the results in a work and a domestic environment, both co-inhabited by robots and humans wearing smartwatches.

2 SYSTEM ARCHITECTURE

The proposed system includes two types of actors: a mobile assistive robot and a mobile device worn by the user, e.g. smartwatch. The information flow among the actors is shown in Figure 1.

We assume that the mobile robot is equipped with proprioceptive sensors, such as wheel encoders or an inertial sensor and a distance sensor such as a laser range finder, sonars, or infrared sensors. Those sensors are required in order for the robot to create a map of the environment and localise therein, as well as perform basic navigation in it. We don't rely on camera sensors, since cameras are very privacy-intrusive and would also pose severe privacy issues in home environments.

We make the assumption that the user is carrying a smart device, e.g. a smartwatch on her right arm if right-handed or left arm if left-handed. Considering that smartwatches have been gaining steadily in public acceptance, our assumption is mild. In fact, smartwatches are a sensible choice for detecting human activities from inertial data, and are not intrusive compared to other sensors.

3 SEMANTIC MAPPING PHASE

In this section we describe the first phase of our approach, in which the robot is able to create a semantic map on top of the metric map of the environment, by accumulating the user activities information over time, during robot-user colocation events. We start by introducing some basics of *Bidirectional Long-Short Term Memory* (BLSTM) neural networks for this work and then describe our proposed activity classification network architecture. Finally, we introduce the semantic mapping creation process.

3.1 Human activity recognition

Long-Short Term Memory (LSTM) networks were introduced as a modified version of *Recurrent Neural Networks* (RNNs), in order to address the vanishing point problem, through the inclusion of gating cells which allow the network to selectively store and forget past memories. They have recently shown promising results when applied to the problem of human activity recognition [6]. *Bidirectional LSTMs* (BLSTMs) [5] are a variant composed by one forward LSTM and one backward LSTM running in reverse on the data and with their features concatenated at the output layer. BLSTMs have been found to perform better when dealing with small datasets.

However, traditional neural networks do not offer a Bayesian probabilistic interpretation of the quality of classification results. In order to estimate the uncertainty surrounding our classification

results we applied the approach of *variational* LSTMs [4] to the problem of activity recognition. In [4] the authors suggested the use of dropout in LSTMSs for approximate Bayesian inference. Dropout is also used in the recurrent connections, and the same dropout masks are repeated at each time step for inputs, outputs, and recurrent layers. Variational LSTMs have been shown to outperform the classic variant, while at the same time offering a useful Bayesian representation of the output. However, to our knowledge they have not yet been explored in the context of human activity recognition.

The input gate $\mathbf{g}^i$ controls how the input enters into the contents of the memory cell for the current time-step. The forget gate, $\mathbf{g}^f$, determines when the memory cell should be emptied by producing a control signal in the range 0 to 1 which clears the memory cell as needed. The output gate $\mathbf{g}^o$ determines whether the contents of the memory cell should be used at the current time-step. $\mathbf{g}^c$ is the cell state vector.

$$\begin{aligned}
\mathbf{g}^i &= \sigma(\mathbf{W}^i * (\mathbf{h}_{t-1} \odot \mathbf{z}_h) + \mathbf{I}^i * (\mathbf{x}_t \odot \mathbf{z}_x)) \\
\mathbf{g}^f &= \sigma(\mathbf{W}^f * (\mathbf{h}_{t-1} \odot \mathbf{z}_h) + \mathbf{I}^f * (\mathbf{x}_t \odot \mathbf{z}_x)) \\
\mathbf{g}^o &= \sigma(\mathbf{W}^o * (\mathbf{h}_{t-1} \odot \mathbf{z}_h) + \mathbf{I}^o * (\mathbf{x}_t \odot \mathbf{z}_x)) \\
\mathbf{g}^c &= \tanh(\mathbf{W}^c * (\mathbf{h}_{t-1} \odot \mathbf{z}_h) + \mathbf{I}^c * (\mathbf{x}_t \odot \mathbf{z}_x)) \\
\mathbf{m}_t &= \mathbf{g}^f \odot \mathbf{m}_{t-1} + \mathbf{g}^i \odot \mathbf{g}^c \\
\mathbf{h}_t &= \tanh(\mathbf{g}^o \odot \mathbf{m}_{t-1})
\end{aligned} \tag{1}$$

where $\mathbf{W}^u$, $\mathbf{W}^f$, $\mathbf{W}^o$, $\mathbf{W}^c$ are weight matrices and $\mathbf{I}^u$, $\mathbf{I}^f$, $\mathbf{I}^o$, $\mathbf{I}^c$ are projection matrices. σ is the logistic sigmoid function. $\mathbf{m}_t$ is the internal state of the cell and $\mathbf{h}_t$ is the hidden vector. $\mathbf{z}_x, \mathbf{z}_h$ are random binary masks that remain constant at each step.

The other difference from standard LSTMs is that at prediction time the dropout remains active. Each prediction is repeated n times, in our case 50 times, and it is possible to compute the mean class prediction and the associated variance over the set of n samples.

3.2 Semantic map inference

Detecting rooms As in [10], at the lower level a SLAM algorithm creates a grid map of the environment using the robot sensors. Using a template-based door detector [10] on laser distance data, the robot is able to group together multiple cells into individual rooms. We use the concept of *room* in a broad sense to denote both regular rooms and corridors. The aim of semantic mapping is to assign semantic categorical labels (e.g. kitchen, bathroom, corridor, etc.) to each cell in the grid.

Detecting colocation events Once the robot builds a grid map of the environment, it starts roaming through it, and records any colocation events with users. For detecting the user's position relative to the robot, we use a leg detector on distance data coming from the robot. In order to ensure that the detected person is indeed the user wearing the smartwatch, we place one BLE emitter on the robot and measure the received signal strength at the smartwatch. On detecting the beacon, the smartwatch sends to the robot the user identifier along with the current detected activity distribution. Whenever the robot detects the user and receives probabilistic activity data from that user, it triggers a colocation event.

Semantic map updates Each cell c in the robot's grid map is assigned a vector $M(c)$ indicating the probability that cell c belongs to a room of a particular type.

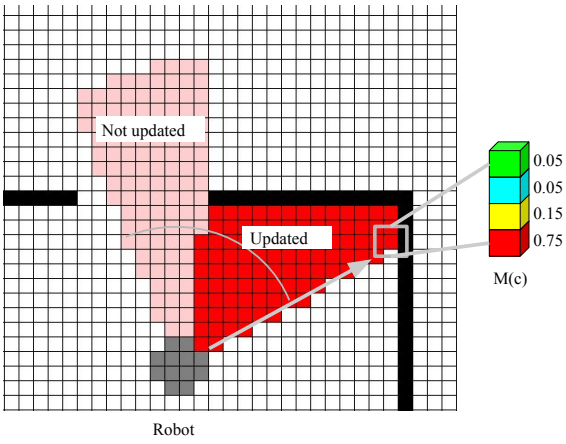


Figure 2: Grid mapping update. Each cell is represented as a vector of room type probabilities (shown in different colors), and is updated based on laser observations using a raytracing procedure along each laser measurement.

On detecting a colocation event, the robot highlights a number of cells that are within its view, with the intention of updating their semantic map probabilities. Figure 2 shows the cells that are within the sensing range of the robot when it detects a person nearby. Note that if a robot is situated in a room and looks in the direction of the door, it ignores those cells that are beyond the door frame.

The probabilities of selected cells having different room types are then updated as follows.

$$M(c)^r := M(c)^r \times \sum_a p(r|a) \times p(a) \tag{2}$$

where $M(c)^r$ is probability of cell c of belonging to room type r, $p(r|a)$ is the probability of being in a room given activity a, and $p(a)$ is the probability that the user is actually performing that activity. In practice this is implemented as a sum of logs of the prior and conditional probabilities, instead of a product of probabilities [13].

Probabilities $p(r|a)$ are drawn from the *Concept net* open source knowledge graph [1], which gives a list of all possible activities associated with each room type, with a weight that represents the strength of the relationship between room and activity. We can exploit these weights, after normalization, in order to obtain usable priors.

The semantic map is updated after each robot-user colocation event.

4 USER LOCALISATION PHASE

In this section we propose a simple graphical model for room-level localisation based on Hidden Markov Models. The model is based on the joint probability distribution between user location and activity. The states of the model represent semantic room types and the transitions represent the transition probability between different room types, e.g. from kitchen to bathroom.

The model alternates between two phases, depending on the predicted activity, namely a *walking phase*, and a *stationary activity*

phase. If a series of walking activities are detected, the model estimates the length of the walking phase in seconds (this is possible since activities are detected at a constant rate) and treats it as a single walking event, representing a transition between two nodes.

Otherwise, if another activity is detected, the model updates the probability distribution of each node according to emission probabilities, as in a classical HMM.

In summary, the walking activity events are concatenated into a single walking event which acts as a control input in the HMM, and impacts the transition probability between different room types.

In the stationary activity phase, state probabilities are only updated using emission probabilities.

5 EXPERIMENTAL RESULTS

The system was implemented using the *Robot Operating System* (ROS) and the Keras library.

5.1 User activities

Data collection protocol For training our network, we gathered inertial data from a set of 20 users of ages between 24 and 60 (with μ=31). Users were given a smartwatch (Sony Smartwatch 3) to be worn on their right hand if right-handed or on the left if left-handed. We defined a list of complex daily activities typical of domestic environments. Each subject was asked to perform the activities, one by one, based on his/her own interpretation and style. In order to sufficiently sample the continuous movement of non-transient actions, each subject was asked to perform each activity continuously for 60 seconds or more.

Two are simple activities (walking, idling), while the rest are complex activities that are typically performed very differently by different people and in different environments (e.g., washing dishes, brushing teeth, using a PC). 3 hours and 10 minutes of data were collected in total.

Training We train our network architecture using standard back-propagation and the ADAM optimizer. For activity recognition the input of the network is a sequence of 3-axial acceleration data and 3-axial angular velocity data of fixed length.

We experimentally found that a window size of 3s offers the best results for complex activity classification in most cases. This is due to the fact that these activities are composed by a series of movements that span over a longer time window, compared to classic activities such as walking, running, biking, etc. We divide the data into windows of 3s with an overlap of 50%. The data is subsampled to a frequency of 50 Hz.

Figure 3 shows the classification results. The network achieves an accuracy of 87.5% on the test set.

5.2 Semantic mapping

We test the semantic mapping in both an office-like environment and a domestic environment. In our experiments, users are equipped with a smartwatch, connected via WiFi to the robot. The robot is a Turtlebot 2 equipped with a Microsoft Kinect camera. The camera is only used to simulate a laser range finder to localize in the map, to detect doors using a template matching algorithm and to detect the user's position during colocations. The first scenario is an office-like environment, composed by a series of rooms and

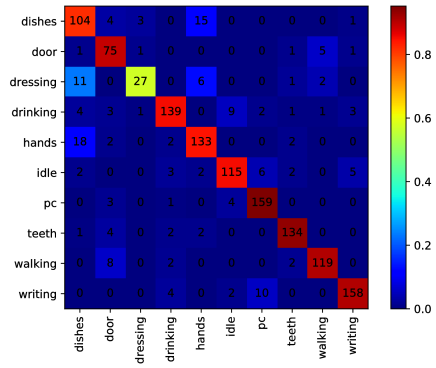
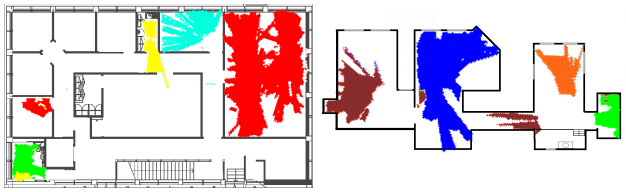


Figure 3: Confusion matrix for the variational BLSTM over 10 classes.



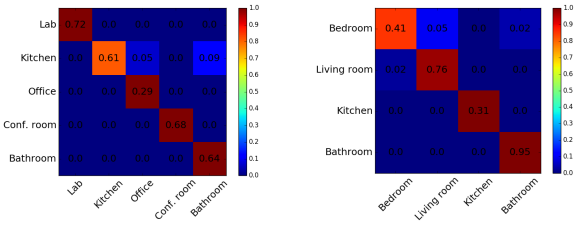
(a) (b)

Figure 4: Resulting semantic maps for the two scenarios. The estimated semantic map is superimposed to a CAD map. Each color correspond to a different room type (red = lab/office, yellow = kitchen, green = bathroom, cyan = conference room, brown = bedroom, blue = living room, orange = dining room).

a corridor. In the first scenario we are interested in mapping five rooms (lab, conference room, kitchen, office, bathroom). There is a sixth multi-purpose room in the center, that is not considered in the experiment. The second is a domestic scenario located inside Keble College in Oxford. In the second scenario we are interested in mapping domestic utilities (bedroom, bathroom, kitchen, dining room, living room).

The experiments lasted for a total of 1 hour per user, with the robot and the user moving in the environment, entering various rooms and triggering colocation episodes. The final semantic maps are shown in Figure 4. The resulting semantic maps are somewhat sparse in certain areas, since there were few colocation episodes. Over a longer period of time we can expect the map to become more complete. Anyway, the results show how the robot is able to compute a relatively dense and accurate semantic map just from a few colocation events, in both office-like and domestic scenarios.

Figure 5 reports the ratio of map cells identified as a particular room type for the five rooms in the two environments. The values are computed as the ratio between the cells classified as a particular room type and the total number of cells in each room. Note that the final mapped area is dependent on the presence of furniture or obstacles and on the trajectory of the robot. The values reflect the fact that only partial areas of each room have been mapped. For



(a) (b)

Figure 5: Confusion matrix for the two scenarios. Each row represents one room; each column represent a semantic label; we report the percentage of cells in each room that are classified with a particular semantic label.

instance as the office was occupied by a desk and several chairs, the robot could not reach the whole room.

5.3 User localisation

In this experiment we show how we can combine the semantic map obtained in the first phase and successive colocation events in order to learn the parameters of a simple graphical model for user localisation at room-level, independently from the robot. We perform these experimental tests in the same two scenarios of the previous experiment. Inertial data was collected from a test set of 5 users. We show the localisation results and compare them with the ground-truth location, which is obtained by placing BLE beacons in each room of interest in both scenarios.

The system first learns the correlation between room locations and activities, in the form of emission probabilities for the different activities given room types, over a series of colocation events over time. Since the robot has access to the semantic maps from the previous experiment, it is able to learn the emission probabilities over time. We expect that the activities performed in rooms which are of the same type to be similar (e.g. lab and office), so in this experiment we combine the two room types. As the classes are somewhat unbalanced (e.g., people tended to spend most of the working day in the lab), the classification accuracy for each specific class is weighted by the number of samples in the class.

The room-to-room distances used to estimate transition probabilities are obtained from the map built by the robot.

We provide the results of the localisation module in Table 1 for both environments on a test set of 5 users. In Figure 6 we show the detected activities along with the predicted room locations for one user in the second scenario, over a window of 30 minutes.

	Precision	Recall	f_1 score
Office-like	0.81	0.91	0.86
Domestic	0.87	0.95	0.91

Table 1: Prediction accuracy of user localisation for both scenarios.

6 CONCLUSIONS

This work presented a framework that integrates assistive robots and consumer wearable devices, for sharing information about room utility between robots and users. In our scenario, the robot and the user coexist in a workplace or household. The robot creates a map using any sensor that can provide distance measurements, then it is able to navigate the environment using standard algorithms. The user wears a smartwatch that continuously acquires inertial data. When the robot and the user are in the same room, meaningful user activities are used to add semantic meaning to the map in the form of room type probability. We then proposed the use of a variational B-LSTM network for recognizing complex spatio-temporal activities from raw data, that keeps the whole framework probabilistic. In a second phase, when a semantic map is available and the robot detects the user, raw data from the user's wearable device can be used to detect room types. We trained a simple graphical model to provide room level localisation for the user even in the absence of the robot. In the model, nodes represent room types and transitions represent transitions between room types. This enables the robot to be aware of the room location of the user at any time.

7 ACKNOWLEDGMENTS

This work has been partly supported by the EPSRC Program Grant "Mobile Robotics: Enabling a Pervasive Technology of the Future (GoW EP/M019918/1)" and by EU's Horizon 2020 program under the Marie Skłodowska-Curie grant No 722022.

REFERENCES

[1] 2018. Conceptnet.io. http://conceptnet.io/ [Online; accessed May-2018].
[2] Abdulrahman Alarifi, AbdulMalik Al-Salman, Mansour Alsaleh, Ahmad Alnafessah, Suheer Al-Hadhrami, Mai Al-Ammar, and Hend Al-Khalifa. 2016. Ultra Wideband Indoor Positioning Technologies: Analysis and Recent Advances. *Sensors* 16, 5 (May 2016), 707. https://doi.org/10.3390/s16050707
[3] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian D Reid, and John J Leonard. 2016. Past, Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age.
[4] Yarin Gal. 2015. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *arXiv:1512.05287* (2015).
[5] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*. Springer, 799–804.
[6] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 1533–1540. http://dl.acm.org/citation.cfm?id=3060832.3060835
[7] Michael Hardegger, Daniel Roggen, Alberto Calatroni, and Gerhard Tröster. 2016. S-SMART: A Unified Bayesian Framework for Simultaneous Semantic Mapping, Activity Recognition, and Tracking. *ACM Trans. Intell. Syst. Technol.* (2016), 28.
[8] Michael Hardegger, Daniel Roggen, and Gerhard Tröster. 2015. 3D ActionSLAM: wearable person tracking in multi-floor environments. *Personal and Ubiquitous Computing* 19, 1 (2015), 123–141. https://doi.org/10.1007/s00779-014-0815-y
[9] Seungwoo Lee, Yungeun Kim, Daye Ahn, Rhan Ha, Kyoungwoo Lee, and Hojung Cha. 2015. Non-obstructive room-level locating system in home environments using activity fingerprints from smartwatch. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*. 939–950. https://doi.org/10.1145/2750858.2804272
[10] Andrzej Pronobis and Patric Jensfelt. 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 3515–3522.
[11] Felipe Barbosa Araújo Ramos, Anne Lorayne, Antonio Alexandre Moura Costa, Reudismam Rolim de Sousa, Hyggo Oliveira de Almeida, and Angelo Perkusich. 2016. Combining Smartphone and Smartwatch Sensor Data in Activity Recognition Approaches: an Experimental Evaluation. In *SEKE*.
[12] Juhi Ranjan and Kamin Whitehouse. 2016. Towards recognizing person-object interactions using a single wrist wearable device. In *Proceedings of the 2016 ACM*

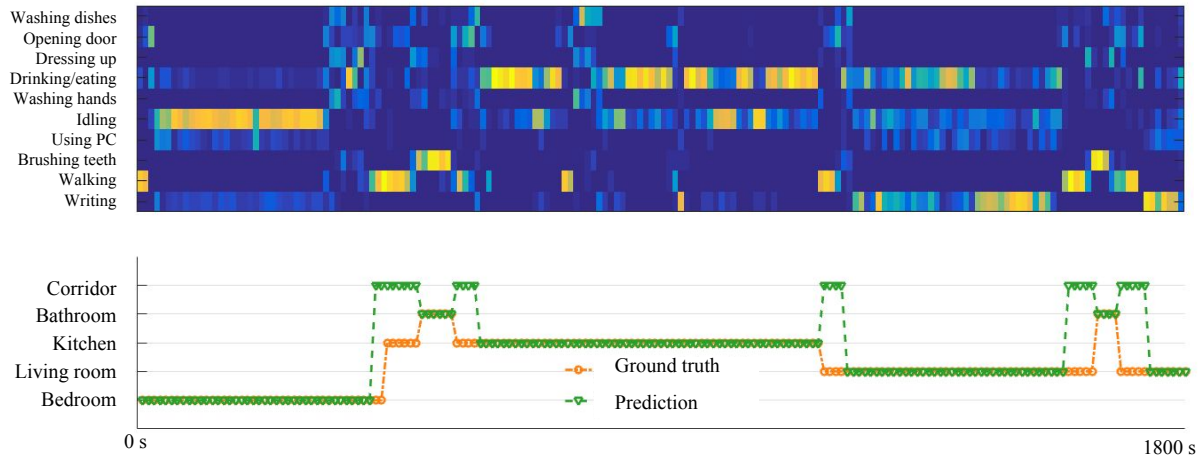


Figure 6: Trace of activities and estimated room location aligned in time. The top image shows the estimated activity probabilities; the bottom image shows the predicted location as well as the ground truth.

International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp Adjunct 2016, Heidelberg, Germany, September 12-16, 2016. 722–731. https://doi.org/10.1145/2968219.2968279

[13] Niko Sunderhauf, Feras Dayoub, Sean McMahon, Ben Talbot, Ruth Schulz, Peter Corke, Gordon Wyeth, Ben Upcroft, and Michael Milford. 2016. Place categorization and semantic mapping on a mobile robot. In *IEEE International Conference on Robotics and Automation (ICRA 2016)*. IEEE, Stockholm, Sweden.

[14] Rui Yao, Guosheng Lin, Qinfeng Shi, and Damith Ranasinghe. 2017. Efficient Dense Labeling of Human Activity Sequences from Wearables using Fully Convolutional Networks. *arXiv preprint arXiv:1702.06212* (2017).

[15] Yuan Zhuang, Jun Yang, You Li, Longning Qi, and Naser El-Sheimy. 2016. Smartphone-based indoor localization with bluetooth low energy beacons. *Sensors* 16, 5 (2016), 596.