STAM: A Framework for Spatio-Temporal Affordance Maps

Francesco Riccio[1(✉)], Roberto Capobianco[1], Marc Hanheide[2], and Daniele Nardi[1]

[1] Department of Computer, Control, and Management Engineering, Sapienza University of Rome, via Ariosto 25, Rome 00185, Italy
{riccio,capobianco,nardi}@dis.uniroma1.it
[2] Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, Brayford Pool, Lincolnshire, Lincoln LN6 7TS, UK
mhanheide@lincoln.ac.uk

Abstract. Affordances have been introduced in literature as action opportunities that objects offer, and used in robotics to semantically represent their interconnection. However, when considering an environment instead of an object, the problem becomes more complex due to the dynamism of its state. To tackle this issue, we introduce the concept of Spatio-Temporal Affordances (STA) and Spatio-Temporal Affordance Map (STAM). Using this formalism, we encode action semantics related to the environment to improve task execution capabilities of an autonomous robot. We experimentally validate our approach to support the execution of robot tasks by showing that affordances encode accurate semantics of the environment.

Keywords: Spatial knowledge · Affordances · Semantic agents

1 Introduction

The concept of affordances has been originally introduced by Gibson [4] as action opportunities that objects offer. This idea has been recently used in robotics to learn [6], represent [11] and exploit [5] object related actions in human-populated environments. However, when considering the affordances of an environment, methods proposed in literature cannot be directly applied. Differently from normal objects, the state of the environment is highly dynamic and contains the state of the robot and other dynamic entities, such as humans. This inevitably leads to a more complex problem that requires specific representation and learning approaches.

To tackle this problem, the concept of spatial affordance has been adopted in some works with the aim of supporting navigation [3] or improving the performance of a tracking system. In this work, we use this concept to encode action semantics related to the environment to improve task execution capabilities of an autonomous robot. In particular, we formalize a Spatio-Temporal Affordance

J. Hodicky (Ed.): MESAS 2016, LNCS 9991, pp. 271–280, 2016.
DOI: 10.1007/978-3-319-47605-6_22

Map (STAM) as a representation that contains high-level semantic properties of an environment, directly grounded on the operational scenario. This grounding is obtained through the use of a function (the affordance function), that generates areas of the environment that afford an action, given a particular state or an equivalent observation of the world. More in detail, STAM contains generic descriptors that (if needed) provide prior information about the actions. For example, when performing a following task, we might not want the relative distance of a robot, with respect to the followed individual, to be greater than a given threshold. These descriptors are then specialized according to the environment where the robot is operating – i.e., the current state of the external world, its entities, including objects and people, and their position over time.

We evaluate an autonomous STAM agent over the execution of a following task that, as shown in Fig. 1, can be beneficial in several applications. In this example, we use expert demonstrations to teach a robot the spatial relation that holds between the environment and the task "to follow". While learning

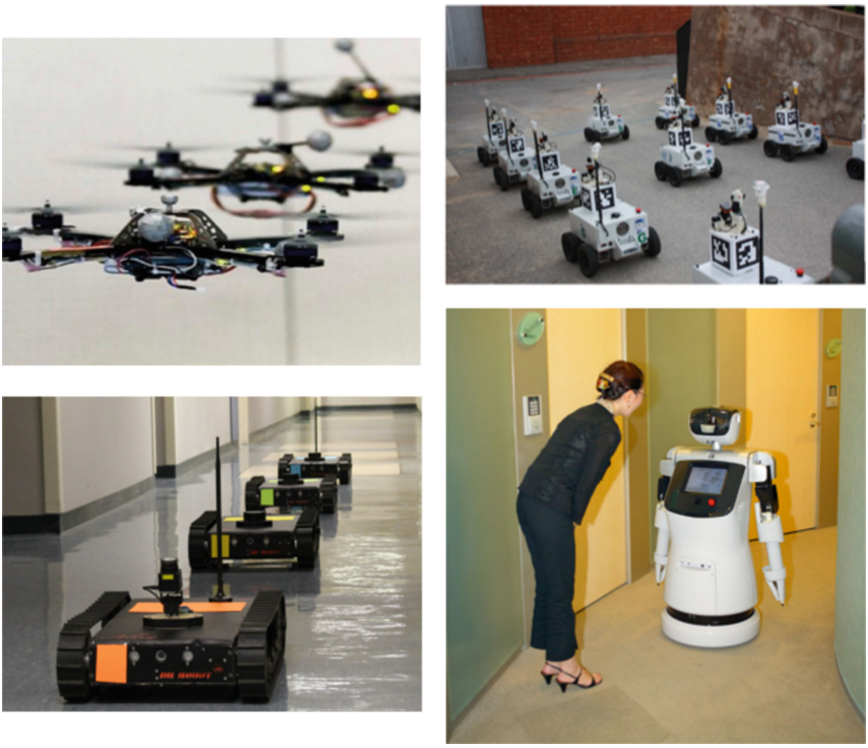


Fig. 1. Following is a key skill in several robotic applications: swarm airdrones, robot teaming, exploration and service tasks. All of them, however, require the robot to execute the task according to different criteria, such as closeness or social acceptability. Being able to represent (and learn) the task semantics according to the specific scenario improves task execution capabilities of a robot.

from demonstration has been already used to learn object affordances [10], we provide an example of how to easily extend such techniques to the case of spatial affordances. Our tests demonstrate that affordances encode accurate semantics of the environment and that they can be used to improve robot skills in terms of efficiency and acceptability in the specific context.

The remainder of this paper is organized as follows. Section 2 presents previous research about affordances, while Sect. 3 defines the concept of Spatio-Temporal Affordances and Spatio-Temporal Affordance Maps, by also describing how they can be generated (Sect. 3.1). Additionally, Sect. 4 describes how to use STAM on a robot and Sect. 5 reports on our experimental validation. Final conclusions are presented in Sect. 6.

2 Related Work

Affordance theory has been introduced to represent possible actions that a robot can perform over a particular object. We extend affordance theory to explicitly formalize the environment itself as a combination of spatial affordances that are used to provide a semantic analysis of the space surrounding the robot. In this context, there is not a vast literature that represent spatial affordances and no prior work models affordances in a general framework. In fact, affordances are used to leverage a particular robot behavior or to adapt the routine of a specific algorithm. For example, Epstein et al. [3] exploit spatial affordances to support navigation. In this work, the leaned spatial affordance informs the robot about the most suitable action to execute for navigation. However, this approach cannot be generalized, since the affordance model strictly depends on a metric representation of the operational scenario. Hence, different representations, such as topological maps, cannot be used. Similarly to our work, Diego et al. [2] encode activities in an affordance map in order to leverage robot movements. The affordance map is used to represent the presence of people in the environment and then to avoid crowded areas not easily navigable. In a different scenario, Luber et al. [9] use affordances to improve tracking and prediction of people destinations. Also in this case, the authors exploit spatial affordances to map activities directly into the operational scenario. However, their system is not intended to run on a robot, and the activities recognized only relate to the presence of people in the scenario. The aforementioned works formalize spatial affordances to only represent navigability of the environment, and in most of the cases, the proposed approaches cannot encode spatial semantics which is a key contribution of our work.

Manifold works confirm our insights that a proper spatial semantic representation can improve robot capabilities. These works typically evaluate spatial semantics although they do not explicitly represent spatial affordances. For example, Rogers et al. [12] and Kunze et al. [7] exploit semantic knowledge to afford a search task. In [12], a robot attaches a semantic label to each room of an environment, and considers the semantic link between the object to search and locations in the indoor scenario. However, the used semantic annotation is very

coarse and remains static once acquired. In [7], the authors compare different areas of the environment depending on flat surfaces and the semantic label of objects previously seen in the scene. Also in this case the proposed framework is instantiated to a particular task and the search is only influenced by objects semantics. We believe that object semantics do not provide a complete environmental knowledge and robot performance can be improved in executing these kind of tasks by integrating information about activities and areas where robot actions are performed.

All the aforementioned contributions exploit spatial affordances to model a unique task and to improve robot skills in performing that specific task. In this work, we want to introduce a general architecture that provides the possibility to model different types of spatial affordances simultaneously. To this end, we consider the remarkable contribution of Lu et al. [8]. The authors propose a layered costmap to encode different features of the environment in order to support navigation. Their architecture enables to formalize each layer independently, which is beneficial in the development of robotic systems. We borrow such paradigm and propose a modular approach in representing affordances. Additionally, we generalize our framework by not forcing our system to only represent navigability tasks. As shown in Sect. 3, we propose a system to semantically annotate the space of the environment in order to support manifold high-level tasks – of which navigability is just an instance.

3 STAM: Spatio-Temporal Affordance Map

Affordances have been originally introduced by Gibson [4] as action opportunities that objects offer, and further explored by Chemero [1] in a more recent work. This notion has been accordingly adopted in robotics to provide a different perspective in representing objects and their related actions. Here, we extend the spatial affordance theory, where the considered "object" is the environment itself, by introducing the idea of spatial semantics and spatio-temporal affordances. Spatial semantics provides a connection between the environment and its operational functionality – e.g., in a surveillance task, areas that are hidden or not entirely covered by fixed sensors present a different "risk semantics". A Spatio-Temporal Affordance (STA) is a function that defines areas of the operational environment that afford an action, given a particular state of the world.

Definition 1. *A spatio-temporal affordance (STA) is a function*

$$f_{E,\boldsymbol{\theta}} : S \times T \rightarrow \mathrm{A}_E. \tag{1}$$

$f_{E,\boldsymbol{\theta}}$ depends on the environment E and a set of parameters $\boldsymbol{\theta}$ characterizing the affordance function. It takes as input the state of the environment $s_E(t) \in S$ at time t, a set of tasks $\{\tau(t)\} \in T$ to be performed, and outputs a map of the environment A_E that evaluates the likelihood of each area of E to afford $\{\tau(t)\}$ in s_E at time t.

The function $f_{E,\boldsymbol{\theta}}$ hence characterizes spatial semantics by evaluating areas of E where the set of tasks $\{\tau(t)\}$ can be afforded. At each time t, it generates the spatial distribution of affordances within the environment and encodes them in a map A_E. Then, the STA function can be exploited by an autonomous agent as a part of a Spatio-Temporal Affordance Map (STAM) - a representation that encodes the semantics of the agent's actions related to the environment.

Definition 2. *A Spatio-Temporal Affordance Map (STAM) is a representation of the STA of an environment that can be (1) learned, (2) updated and (3) used by an autonomous agent to modify its own behavior.*

As depicted in Fig. 2, the core element of a STAM is the function $f_{E,\boldsymbol{\theta}}$ introduced in Definition 1, that depends on a set of parameters $\boldsymbol{\theta}$ obtained from an *affordance description module* and takes as input the current state of the world and a set of tasks from the *environment module*. In particular:

- the affordance description module (**a-module**) is a knowledge base composed by a library of parameters $\boldsymbol{\theta}$ that characterize the STA and represent its *signature*. The signature modifies the spatial distribution of affordances within the environment;
- the *environment module* (**e-module**) encodes the state of the world $s_E(t)$ and provides such a state to the STA function, by coupling it with a set of tasks $\{\tau(t)\}$ to be executed in order to achieve the desired goal.

It is worth remarking that $f_{E,\boldsymbol{\theta}}$, s_E and A_E refer to a common representation of the environment E that needs to be instantiated in order to enable a robot to use STAM. Such a representation can be chosen to be a metric map, a grid map, a topological map, or a semantic map. Additionally, STAM can be used to interpret relations among different affordances (if there exists) and to represent affordances individually. In fact, as shown in Fig. 3, a STA can be seen as a

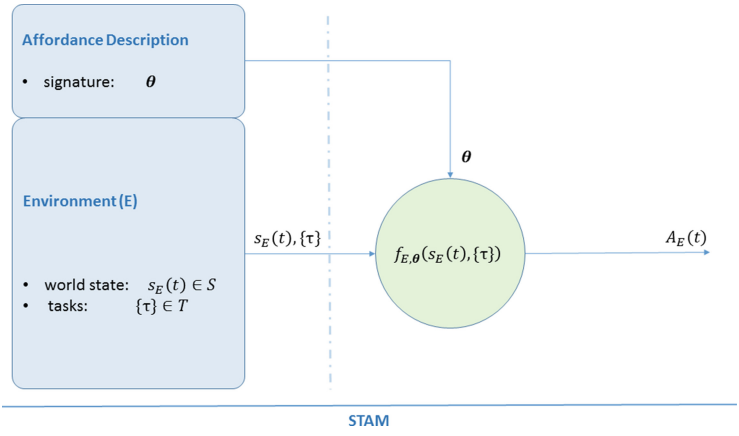


Fig. 2. Spatio-temporal affordance map – STAM.

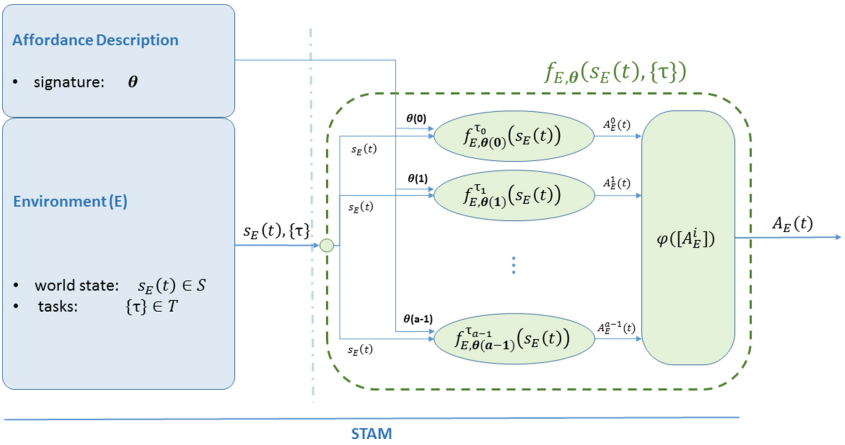


Fig. 3. Spatio-temporal affordance map – STAM.

composition of different $f^{\tau_i}_{E,\boldsymbol{\theta}(i)}$ functions ($i \in [0, a-1]$, where a is the number of affordances), each modeling the spatial distribution A^i_E of a particular affordance in E. These are then combined by a function ϕ, that takes as input all the A^i_E and outputs a map A_E that satisfies $\{\tau(t)\}$, according to the considered affordances.

3.1 Generating a Spatio-Temporal Affordance Map

The affordance map A_E is a representation of the operational environment that evaluates E with respect to the current state of the world and encodes areas of E where a particular task can be afforded. For instance, in the case in which the environment is represented as a grid-map, A_E encodes in each cell the likelihood of a given area to afford an action. According to Definition 1, the generation of A_E directly depends on a general set of parameters $\boldsymbol{\theta}$ – the affordance signature – that modify how affordances model the space. Hence, they constitute the main vehicle to shape affordances and need to be carefully designed or learned. In the first case, accurate understanding of each parameter $\boldsymbol{\theta}$ and the function $f_{E,\boldsymbol{\theta}}$ is required. In the latter case, the STA function can be implemented as regression or classification algorithm, and standard gradient-based methods can be used to update $\boldsymbol{\theta}$. For instance, when learning affordances from observations of other agents' behaviors (e.g., humans) a neural network could be used. In this case, the set of parameters $\boldsymbol{\theta}$ would represent the connection weights between different layers and they could be computed by means of back-propagation.

4 Using STAM on a Robot

STAM is intended to directly influence the behavior of an autonomous agent and, in particular, the navigation stack of a mobile robot. We consider the case

in which the robot navigation system relies upon standard costmap-based techniques [8]. In contrast to previous work in this field, we are not interested in enabling a robot to "*go from point A to point B*", but we aim at making the agent capable to "*go from A to* β", where $\beta \in B'$ is a set of "good" poses obtained from the map A_E generated by STAM. Such poses intrinsically respect spatio-temporal constrains imposed by the considered affordances. Among these, the selection of the final pose can be based upon different criteria, such as the top scoring area in A_E, the nearest area to the robot, the biggest area, or a combination of these criteria. Nevertheless, we also want the robot to decide how to navigate the environment by selecting the path accordingly to the affordances imposed by the task. To this end, we can directly use A_E to effectively crop out all the trajectories of the robot that cross areas violating affordance constraints. In particular, we can substitute the costmap with a *gainmap* that encodes high-level information extracted from STAM. Accordingly, the robot will not follow the cheapest path, as in "usual" costmap-based systems, but it will maximize its gain over the generated gainmap. Such a map is generated as a function of the normalized cost map and the likelihood obtained from A_E.

$$m(cost, likelihood, \lambda) = \lambda(1 - cost) + (1 - \lambda)likelihood, \tag{2}$$

with $\lambda \in [0, 1]$. In this respect, we are modifying the navigation systems of an autonomous robot by transferring high-level information encoded in A_E into the navigation system.

5 Experiments

In order to evaluate of our approach we perform an analysis of the learned affordance model. To this end, we exploit expert demonstrations to teach a robot how to correctly interpret the environment when performing a following task. Then, we evaluate the learned model by reporting the affordance map A_E generated by the affordance function and the prediction error of the regression algorithm after each demonstration.

5.1 Affordance of a Following Task

We consider a robot that has to perform a following task. In this case, the areas of the environment E that afford the task depend on manifold factors such as general rules (e.g., forbidden areas), user preferences (that can be encoded in the set of parameters $\boldsymbol{\theta}$) and the position of the followed person (encoded in the state of the environment s_E). According to Definition 1, we can generate A_E and identify robot poses that support the execution of the task. To this end, we encode the pose $\langle x_T, y_T, \alpha_T \rangle$ of the target T to follow in the state $s_E(t)$. Additionally, we use Gaussian Mixture Models (GMMs) and Gaussian Mixture Regression to represent and implement the function $f_{E,\boldsymbol{\theta}}$. The signature $\boldsymbol{\theta}$ of the STA function is hence composed as a tuple $\boldsymbol{\theta} = \langle \pi_1, \mu_1, \Sigma_1, \ldots, \pi_N, \mu_N, \Sigma_N \rangle$,

where π_i is the prior, μ_i the mean vector and Σ_i the covariance matrix of a mixture of N Gaussians.

In this experiment, the signature $\boldsymbol{\theta}$ is learned from demonstration of different experts. To collect expert data we setup two robots in a simulated environment – one randomly navigates, the other is controlled by an expert through a joystick and follows the target robot T by always moving between a minimum and maximum distance from it. During these sessions, the state $s_E(t)$, as defined above, is recorded at each time instant together with the pose $\langle x_F, y_F, \alpha_F \rangle$ of the follower F. The collected measurements are provided as input to the GMM and, by using Expectation Maximization, the tuple $\boldsymbol{\theta} = \langle \pi_1, \mu_1, \Sigma_1, \ldots, \pi_N, \mu_N, \Sigma_N \rangle$ that best fits the data is determined. In our experiments, prior to Expectation Maximization, the model has been initialized with k-means and a set of candidate GMMs has been computed with up to 8 components; the number of components has then been selected to minimize the Bayesian Information Criterion.

The learned model is used by the follower to determine, through Gaussian Mixture Regression, areas of E that enable the robot to execute the task and, hence, to generate A_E. In particular, the output of the regression consists of a mean vector and covariance matrix that enable us to infer the probability distribution (shown in Fig. 4) of the follower pose, given the target pose for the following task τ. In this example, no specific constraint is imposed to the robot for the selection of its path. Hence, the agent can select the pose that

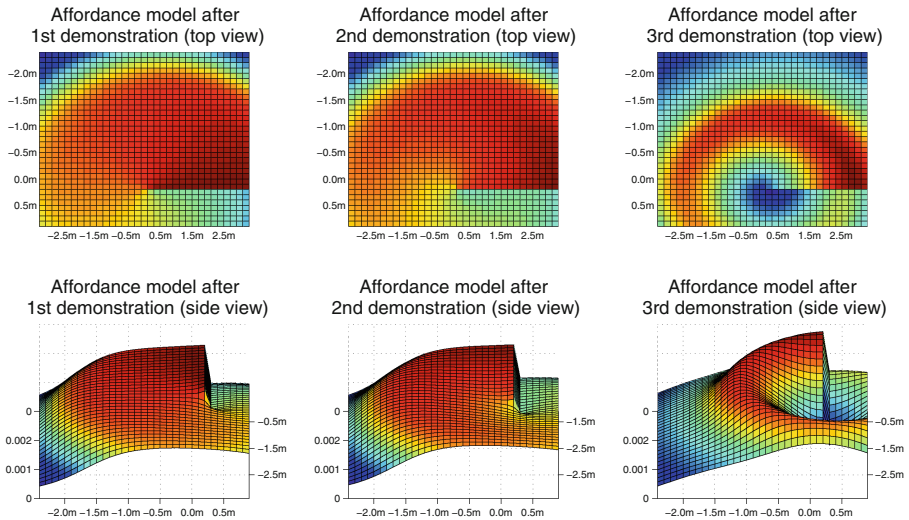
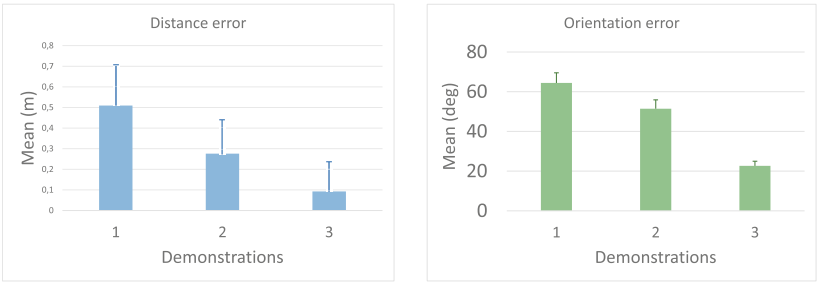


Fig. 4. Spatio-temporal affordance of a following task learned with increasing number of expert demonstrations. Here, the target is located at the origin and the plots represent the probability density function of a pose to afford the task. The plots, whose coordinates are expressed in meters, show that the model is able to represent both minimum and maximum distances from the target, in accordance with the data provided as demonstrations.



(a) (b)

Fig. 5. Error of the best pose, selected according to the learned model, against the expert behavior. On the left we report (a) the mean and standard deviation of relative distance error between the follower and the target, while on the right (b) the mean and standard deviation of the relative orientation error are shown. These values have been obtained by running 20 experiments and incrementally using three expert demonstrations (arranged on the x-axis).

maximizes its profits over the gainmap computed according to Eq. 2, and reach it by following the shortest path.

Finally, we report an analysis of the prediction error of the affordance model generated by the regression algorithm. To this end, we use expert data collected in three different demonstrations in an incremental fashion – after each demonstration we append new training examples to the previous dataset. Then, we generate the affordance model by splitting the dataset into two distinct parts. One is used to learn the affordance model, while the other is used to compute the error of the best pose, selected according to the learned model, against the expert behavior (the ground-truth). To evaluate our model, we ran the experiment 20 times. Accordingly, Fig. 5 shows the mean and standard deviation of the prediction errors of the relative distance (a) and orientation (b) between the target and the follower position. It is worth remarking that, as soon as the affordance model becomes more accurate (Fig. 4), the prediction errors of both the distance and orientation decay.

6 Conclusion

In this paper we presented and formalized Spatio-Temporal Affordances (STA) and Spatio-Temporal Affordance Maps (STAM) as a novel framework to represent spatial semantics. This is a relevant problem since, by providing a connection between the environment and its operational functionality, spatial semantics leads to a proper interpretation of the environment and hence to a better execution of robot tasks. To test this representation, we implemented STAM and learned the affordance model of a following task by exploiting expert demonstrations. Specifically, we set up a simulated environment where human experts

could teach the robot how to correctly interpret the environment when performing a following task. After training, we let our system infer the best position to be in order to follow a target. Results show that (1) the mapping between the space and its affordance is qualitatively valid and (2) the error generated by the use of our model decreases when it becomes more accurate, through the use of a larger number of expert demonstrations.

Nevertheless, learning the affordance of a following task is only a simple and specific use case of STAM. For this reason, in future work we aim at using STAM to run different experiments with manifold tasks and, specifically, to enable a robot to interpret spatial semantics to improve human-robot interactions.

References

1. Chemero, A.: An outline of a theory of affordances. Ecol. Psychol. **15**(2), 181–195 (2003)
2. Diego, G., Arras, T.K.O.: Please do not disturb! Minimum interference coverage for social robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1968–1973, September 2011
3. Epstein, S.L., Aroor, A., Evanusa, M., Sklar, E., Parsons, S.: Navigation with learned spatial affordances. In: COGSCI (2015)
4. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
5. Kim, D.I., Sukhatme, G.S.: Interactive affordance map building for a robotic task. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4581–4586. IEEE (2015)
6. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. The Int. J. Robot. Res. **32**(8), 951–970 (2013)
7. Kunze, L., Burbridge, C., Hawes, N.: Bootstrapping probabilistic models of qualitative spatial relations for active visual object search. In: AAAI Spring Symposium on Qualitative Representations for Robots. Stanford University in Palo Alto, California, 24–26 March 2014
8. Lu, D.V., Hershberger, D., Smart, W.D.: Layered costmaps for context-sensitive navigation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 709–715, September 2014
9. Luber, M., Tipaldi, G.D., Arras, K.O.: Place-dependent people tracking. In: Pradalier, C., Siegwart, R., Hirzinger, G. (eds.) Robotics Research. STAR, vol. 70, pp. 557–572. Springer, Heidelberg (2011)
10. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: from sensory-motor coordination to imitation. IEEE Trans. Robot. **24**(1), 15–26 (2008)
11. Pandey, A.K., Alami, R.: Taskability graph: towards analyzing effort based agent-agent affordances. In: 2012 IEEE RO-MAN, pp. 791–796. IEEE (2012)
12. Rogers, J.G., Christensen, H.I.: Robot planning with a semantic map. In: 2013 IEEE International Conference on Robotics and Automation (ICRA), pp. 2239–2244, May 2013