Probabilistic Spatio-Temporal Fusion of Affordances for Grasping and Manipulation

Christoph Pohl and Tamim Asfour

***Abstract*—Robust vision-based grasping and manipulation of unknown objects in unstructured scenes requires the extraction of action candidates based on visual information while taking into account noise and occlusions in such scenes. We address this problem by combining the concept of affordances and *Bayesian Recursive State Estimation*. We propose to extract affordances using heuristics on the averaged local surface information of supervoxels in a point cloud. Based on a local, geometry-aware coordinate frame, we define a uniform state for different affordances. Using Bayesian statistics, this state is fused across multiple observations of the scene to improve the estimates for the pose and existence certainty of actions. This facilitates the extraction of robust grasping and manipulation actions independent of the segmentation of a scene. The proposed approach is evaluated in grasping experiments with more than 900 grasp executions using the humanoid robot ARMAR-6 in an unstructured scene with a variable number of unknown objects. The experimental results show that the grasping success rate is improved by over 10% compared to a state-of-the-art approach.**

***Index Terms*—Perception for Grasping and Manipulation, Semantic Scene Understanding, Probabilistic Inference**

I. Introduction

THE interaction of an autonomous robot with unstructured and unknown environments based on visual information is still a difficult task. It requires an interpretation of the scene to allow the selection of proper actions that can be executed in a given situation. The ability to interact with cluttered scenes is necessary to increase the robot's autonomy for real-world applications e. g., in inhospitable environments, such as disaster response scenarios or work in contaminated areas, where robots can increase the safety and working conditions for humans. The concept of affordances [1], adapted from cognitive psychology, has been recently employed to endow robots with the ability to extract interaction possibilities and hypotheses for potential actions in the scene based on visual perception. Existing approaches for affordance-based grasping and manipulation rely on the segmentation of the scene, which is often noisy and inaccurate, thus leading to infeasible actions. Additionally, having a measure of the uncertainty of e. g., a grasping pose can be used as an indicator of the grasp's success. For example, in the case of a grasp pose with high uncertainty, the robot could first explore the scene from a different viewpoint before executing such a grasp to increase the probability of success. To deal with uncertainties in an affordance-based extraction of potential actions in a given scene, we present a probabilistic approach to estimate the pose of an action hypothesis together with the certainty of its existence, derived over multiple observations.

Manuscript received: September, 9, 2021; Revised December, 7, 2021; Accepted January, 10, 2022.

This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments. The research leading to these results has received funding from the German Federal Ministry of Education and Research (BMBF) under the competence center ROBDEKON (13N14678).

The Authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. {pohl, asfour}@kit.edu

Digital Object Identifier (DOI): see top of this page.

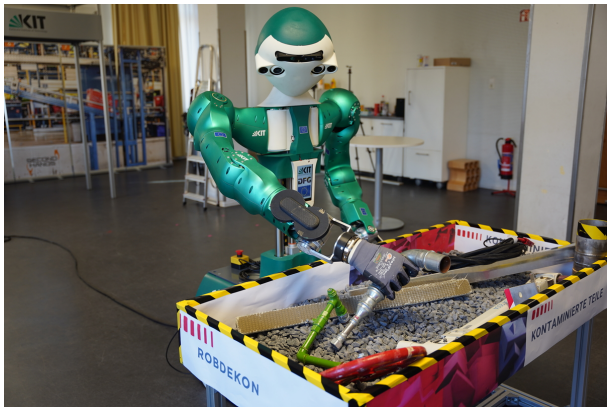


Fig. 1: ARMAR-6 executing an affordance-based grasp.

A. Previous Work

The proposed method builds on our previous work on the computational formalization, extraction, and validation of scene affordances [2], [3], in which we describe affordances as *Dempster-Shafer* belief over the space of end-effector poses. This formulation allows for a hierarchical definition of affordances and the fusion of information from different input modalities. Affordances are defined on primitive shapes, like cylinders, spheres, and planes, and *affordance belief functions* are used to describe the degree of certainty in the existence of an affordance for an end-effector pose. This formulation facilitates the consistent fusion of affordance-related evidence for their validation through physical interaction. However, the representation of affordances as belief over shape surfaces cannot take previous observations for an improved estimate of the end-effector poses into account. Therefore, an entirely different approach to affordances – independent of the representation as primitives shapes – is needed. Estimating the state of a system from multiple observations is a well-understood problem in robotics and can be solved, e. g., using *Recursive State Estimation* [4]. Estimating the 6-dimensional pose remains difficult nevertheless, as conventional recursive

filters in Euclidean space cannot easily handle orientations. However, ready-to-use algorithms for the fusion of poses have recently been developed based on *Bayesian State Estimation* [5].

Therefore, our goal is the probabilistic state estimation of action hypotheses based on the spatio-temporal fusion of visually-extracted affordances from multiple observations. Contrary to our prior work, quadric approximations of the local surface of a point cloud are used to extract curvature information, which is then used to heuristically derive affordances for surface patches that conform to object boundaries. These affordances can be used to generate different types of actions associated with a frame at each surface patch that is uniquely defined through the local surface structure. *Recursive State Estimation* with an *unscented Kalman filter* is then combined with a *Hidden Markov Model* to estimate the *existence certainty* of an action hypothesis and improve its associated pose over multiple observations.

B. Contributions

In this work, we present a *Bayesian Recursive State Estimation* approach for combining multiple observations of affordances to estimate their existence certainty as well as the pose of their associated action's coordinate frame. We evaluate the precision of our approach and its robustness in multiple real-world grasping experiments in a scene with different degrees of clutter on the humanoid robot ARMAR-6 [6]. The major contributions of this work are (1) the probabilistic, spatio-temporal fusion of action-related observations combining an *unscented Kalman filter* and a *Hidden Markov Model*, which is made possible by (2) the definition of a state for different affordances, facilitating the joint fusion of all affordances, while still allowing for versatile action generation.

II. Related Work

Robotic manipulation in real-world scenarios remains a challenging problem, as the robot has to cope with cluttered environments, perceptual uncertainties, and dynamic change of the scene. To identify suitable interaction possibilities for autonomous manipulation in robotics, the concept of affordances [1] was adopted from cognitive psychology. In an affordance-based interpretation of the scene, objects are assigned their respective action opportunities as properties, so that, e. g., a cup might have the "*graspability*" and "*fillability*" affordances. In our previous work [2], [7], affordances have been investigated for the discovery and execution of autonomous actions with a humanoid robot in real-world environments. However, our previous work does not consider the fusion of multiple observations to improve extraction of scene affordances. In [8], affordances derived from local geometric features of object parts are used to define *Conceptual Equivalence Classes*, which state that objects can be treated interchangeably in action execution if they possess parts with the same affordances. While still being dependent on an instance segmentation of the object, the authors of [9] assign affordances to certain keypoints – and not to the entire object – on object classes and show that their system can handle large intra-class variability in a pick-and-place task. There also exist multiple works similar to our approach that treat affordances in a probabilistic framework. In our previous work in [3], affordances were formalized as *Dempster-Shafer* belief functions to facilitate the fusion of belief from different input modalities. A *Markov Logic Network* was used in [10] to extract a probability distribution over grasp affordances of an object to predict the most probable location of a successful grasping action. In [11], a relational affordance model for high-level planning is introduced, which incorporates the current state of the world in terms of random variables and considers object relations between multiple objects. The authors model affordances as a joint probability distribution over objects, actions, and their effects. To the best of our knowledge, there does not yet exist an approach that fuses multiple observations of scene affordances to improve the estimate of the pose, as well as the existence certainty of manipulation actions.

As noisy point clouds can affect the accuracy of vision-based manipulation, surface approximation techniques are employed in many robotic applications. Quadrics [12] are a mathematical construct that can be used to alleviate some of the restraints when working with RGB-D sensors. In [13], quadric surface patches are used for the calculation of stable grasp poses for known objects, which are decomposed to an approximate quadric representation. The approach proposed in [14] is similar to ours, as it uses curved surface patches with a uniquely defined pose based on its curvature to extract a stable foot placement for a bipedal robot. However, in this work, we define one universal parameterization of surface patches for all shapes and affordances. Moreover, we don't rely on a semantic segmentation of the scene, which improves the performance in very cluttered environments.

Recursive State Estimation in robotics [4] has been thoroughly investigated for applications like pose estimation, object tracking, localization, and mapping [15]. Nevertheless, the nonlinear nature of orientations is a problem for traditional approaches like the *Kalman filter*. The statistical treatment of 6D poses as a Lie group [16] has gained attention for state estimation in recent years. In [17], nonlinear filtering based on the unscented transform of dual quaternions is used for a visual SLAM system. For nonlinear state estimation on 6D poses, adaptions to the *unscented Kalman filter* (UKF) [18] show promising results. In [19], the standard UKF is extended for the estimation and modeling of the pose of a quadrotor platform in $\mathrm{SE}(3)$. For the application in general Lie groups, a UKF with partial or full state measurements is proposed in [5]. The authors use concentrated Gaussians on the manifold and define the sigma points in the Lie algebra, therefore removing the need to map them back onto the manifold. The authors in [20] build upon their work and define the time propagation in the Lie algebra as well, instead of on the manifold. This increases the computational efficiency and facilitates the computation of the mean and covariance in the Lie algebra. A Matrix Fisher Distribution directly defined on the manifold $\mathrm{SO}(3)$ is used in [21] instead of Gaussians to implement a UKF for attitude estimation.

III. Approach

An overview of our approach for the extraction and probabilistic fusion of affordance-based actions is given in Figure 2. The approach can be split into three main steps. First, the local surface geometry of a point cloud is analyzed and normals, as well as principal curvatures, are calculated at every point. Based on this, locally consistent surface patches are extracted via an adapted supervoxel clustering (Section III-A). Afterwards, the averaged geometrical features of the supervoxels are used to heuristically define affordances and a temporally consistent coordinate system – the *Local Curvature Frame* – for each patch (Section III-B). In the last step, correspondence likelihoods to previously observed actions are determined and the state of an action hypothesis is updated accordingly using an *unscented Kalman filter* (UKF) and a *Hidden Markov Model* (HMM) (Section III-C). The steps of the approach for the *graspability* affordance are shown in Figure 3

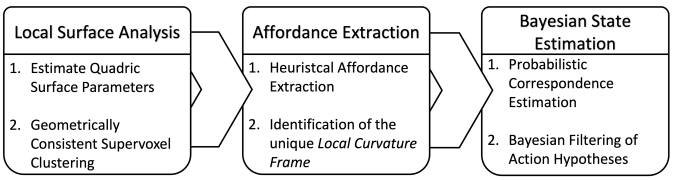


Fig. 2: Overview of the approach.

A. Local Surface Analysis

The first step of the approach is to estimate and extract local surface information from the raw point cloud data, as this is the basis for subsequent processing steps.

1) Quadric Patch Estimation: As a geometrical representation of surfaces in a point cloud, quadrics [12] are used. Quadrics are D-dimensional hypersurfaces embedded in a space of the dimension $(D+1)$, where in this case $D = 2$. By treating surfaces as functions, methods of differential geometry allow the closed-form calculation of important metrics for the local surface structure [22, Chapter 3].

For the approximation of the local surface of a raw point cloud as quadrics, the GPU-implementation described in [23] is used. As a result, the surface normal $\mathbf{n}$ and the principal curvature coefficients $\kappa_{\pm}$, as well as the quadric parameters L, M, N, are obtained. L, M, N are called second fundamental form coefficients and are defined through the *2nd fundamental form* [22].

For the extraction of a coordinate frame defined by the local surface structure in the later sections, the principal directions $\lambda_{\pm}$ are required. They can be obtained from the definition of the curvature through the second fundamental form coefficients:

$$\lambda_{\pm} = -\frac{M}{N - \kappa_{\pm}} = -\frac{L - \kappa_{\pm}}{M}$$

2) Supervoxel Clustering augmented by Local Surface Curvature: After the calculation of the local surface information for each point of the point cloud, the next step is to find regions which have similar properties and cluster them. This can be achieved using different clustering algorithms. We assume one single action hypothesis per region for the manipulation of unknown objects. Therefore, the clustering algorithm described in [24] was chosen. The advantage of this approach is that it adheres to object boundaries and provides an over-segmentation of the scene in the form of supervoxels. To improve the segmentation accuracy, the implementation in the *Point Cloud Library* [25] was adapted to also consider the previously extracted local surface metrics. It is reasonable to assume that points with similar local surface properties, such as the direction of maximum curvature, belong to one semantic segment and, therefore, share affordances (see e. g., [14], [26]).

A supervoxel $V = (\mathbf{t}, \mathbf{c}, \mathbf{n}, \mathbf{k}_{\lambda_-}, K)$ represents a cluster of points and is defined using the averaged features of all points belonging to it, where $\mathbf{t} \in \mathbb{R}^3$ is the position, $\mathbf{c} \in [0 \dots 255]^3$ the color, $\mathbf{n} \in \mathbb{R}^3$ the surface normal, $\mathbf{k}_{\lambda_-} \in \mathbb{R}^3$ the direction of minimal curvature, and $K = \kappa_+ \cdot \kappa_-$ is the Gaussian curvature. Beginning with an initial seeding, the supervoxels are iteratively grown based on the distance d_{vccs} between two adjacent voxels V_1 and V_2 in the feature space given by

$$d_{vccs} = \alpha||\mathbf{t}_2 - \mathbf{t}_1|| + \beta||\mathbf{c}_2 - \mathbf{c}_1|| + \gamma(1 - |\mathbf{n}_1 \cdot \mathbf{n}_2|),$$

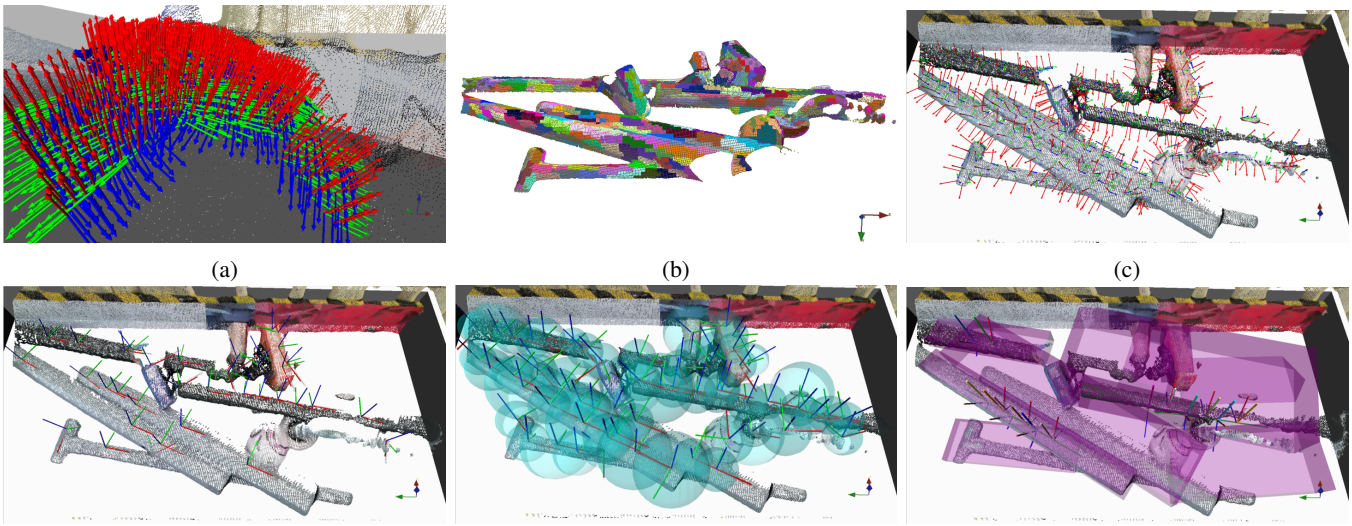
where α, β, γ are scaling constants. To better account for the local surface structure, the feature space was extended by the principal directions and curvatures of the point cloud. Therefore, the new distance in feature space is

$$d_{aug} = d_{vccs} + \delta(1 - |\mathbf{k}_{\lambda_-,1} \cdot \mathbf{k}_{\lambda_-,2}|) \cdot |K_2 - K_1|.$$

B. Affordance Extraction

In our previous work, we introduced affordance-based approaches for the extraction of potential actions for autonomous and semi-autonomous manipulation tasks performed by humanoid robots in unknown environments [7] and presented methods for the probabilistic formalization of affordances [3]. The approach presented in this paper on the other hand, needs a universal state for all affordances to be able to deal with the recursive spatio-temporal fusion of actions.

1) Heuristic Affordance Extraction for Clustered Surface Patches: In our previous work [2], threshold-based decision functions were used to define affordances on extracted environmental geometric primitives. Similarly, heuristics resulting from thresholds of the averaged supervoxel statistics of the local surface geometry are used in this work to extract affordances in the scene. For example, it is natural to assume, that only flat surfaces with a normal that is anti-parallel to the direction of gravity have the affordance of *placability* or *supportability*. Further, we assume that only convex objects (i. e., using the convention of the curvature direction as used in [22] with $\kappa_- \leq 0$) afford *graspability*. An overview of surface metrics used for affordance extraction is given in Table I. Based on these metrics, every point belonging to a supervoxel can be assigned the affordance for which its supervoxel fulfills the requirements.



(a) (b) (c)

(d) (e) (f)

Fig. 3: In (a) the normals, minimal and maximal curvature directions for each point in a surface patch are displayed. (b) shows the clustered supervoxels and (c) the averaged surface information. (d) shows the extracted action observations, and (e) the final, filtered action hypotheses and sphere with a radius equal to $3\sigma_m$. For comparison, (f) shows bounding boxes for the objects based on a segmentation of the scene.

Surface Metric/ Affordance	***Graspability***	***Pushability***	***Placability***
Mean Surface Curvature	Convex	–	Flat
Mean Normal Direction	Upper Hemisphere	Horizontal	Upwards
Volume	$<$ Grasp Volume	–	$>$ Object/ Threshold

TABLE I: Surface metrics for the definition of affordances.

2) Action Generation in a Uniquely Defined, Local Coordinate System: Depending on the affordances extracted for a supervoxel, actions are generated. As $n \geq 0$ affordances can exist for a single supervoxel, i. e., multiple potential actions can be assigned to one supervoxel, a state definition for all action types is needed for the probabilistic state estimation. Otherwise, the temporal fusion would require tracking the state for each affordance, which can quickly become computationally intensive for larger scenes.

For differential surfaces, it holds that the normal direction $\mathbf{n}$ and the principal directions $\mathbf{k}_{\lambda_\pm}$ are always orthogonal [22]. Therefore, we can define a local, geometrically motivated frame $\mathbf{X}$ for each supervoxel, referred to as the *Local Curvature Frame*, that only depends on the averaged curvature and normal of all points belonging to the supervoxel. The frame is chosen in a way that its z-axis aligns with the averaged normal $\bar{\mathbf{n}}$ and its y-axis aligns with the averaged minimal curvature direction of the supervoxel. The execution instructions for the actions can then be generated based on the pose of this *Local Curvature Frame*. For example, a grasp can be generated in a way that the fingers of the hand align with the minimal curvature direction (i. e., y-axis of the *Local Curvature Frame*).

C. Bayesian State Estimation

As the *Local Curvature Frame* for a surface patch is uniquely defined and is associated with the affordances of this patch, it can be used to construct an action's state that facilitates spatio-temporal fusion. Considering that action possibilities in a scene can appear and disappear at any time (e. g., when an object is removed), it is not sufficient to model only the action's pose and its uncertainty, as an additional measure for the *existence* of an action is needed. Therefore, we use a combination of a *Kalman filter* with a *Hidden Markov Model* for the complete probabilistic state estimation of the actions.

We define an action observation $\mathbf{A}_t$, which is linked to the *Local Curvature Frame* with pose $\mathbf{X} \in \mathbb{R}^3 \times \mathrm{SO}(3)$ at time step $t \in \mathbb{R}^+$ and is associated with n affordances a_i

$$\mathbf{A}_t = (\mathbf{X}, t, \{a_1, \ldots, a_n\}).$$

An action hypothesis $\bar{\mathbf{A}}_t$ is the result of combining multiple observations $\mathbf{A}_t$ of an action using the UKF and the HMM

$$\bar{\mathbf{A}}_t = (\bar{\mathbf{X}}, \mathbf{\Sigma}_\mathbf{X}, t, \{p_E^{a_1}, \ldots, p_E^{a_m}\}),$$

where $p_E^{a_i}$ is the *existence certainty* of $\bar{\mathbf{A}}_t$ for the i-th affordance a_i and $\mathbf{\Sigma}_\mathbf{X} \in \mathbb{R}^{6\times6}$ is the covariance matrix of the filtered pose $\bar{\mathbf{X}} \in \mathbb{R}^3 \times \mathrm{SO}(3)$ with the time of last observation t.

For the probabilistic estimation of the state of $\bar{\mathbf{A}}$, two main problems arise: Firstly, a matching hypothesis $\bar{\mathbf{A}}$ needs to be identified for each new action observation, i. e., a hypothesis $\bar{\mathbf{A}}_{t-1}$ corresponding to the currently observed action $\mathbf{A}_t$ needs to be found. Secondly, the current observation $\mathbf{A}_t$ and the previous state estimation $\bar{\mathbf{A}}_{t-1}$ need to be fused to estimate the new state $\bar{\mathbf{A}}_t$ of the action hypothesis.

1) Search for Correspondences between Observations and Hypotheses: For the identification of correspondences between actions in subsequent observations of the scene, we apply a similar approach as for the belief fusion in our previous work [3]. The basic assumption is, that observations of the pose $\mathbf{X}$ of an action hypothesis $\bar{\mathbf{A}}$ are distributed in a Gaussian manner around the mean pose $\bar{\mathbf{X}}$. For the positional part of the pose $\mathbf{t}$, this can be expressed by a multivariate Gaussian *Probability Density Function* (PDF) [4], given that the observation and hypothesis correspond to one another (i. e., conditioned on C)

$$p(\mathbf{t}|C) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_\mathbf{t})}} \exp\left(-\frac{1}{2}(\mathbf{t}-\bar{\mathbf{t}})^T\boldsymbol{\Sigma}_\mathbf{t}^{-1}(\mathbf{t}-\bar{\mathbf{t}})\right),$$

where $\bar{\mathbf{t}} \in \mathbb{R}^3$ is the mean position of the action and $\boldsymbol{\Sigma}_\mathbf{t} \in \mathbb{R}^{3\times3}$ is its covariance matrix.

Since the orientation $\mathbf{R}$ of $\mathbf{A}$ is an element of the special orthogonal group $\mathrm{SO}(3)$, the above probability density function cannot be used without adaption. Following [16], we use local perturbations $\boldsymbol{\tau}$ on the Lie group $\mathrm{SO}(3)$ to model the uncertainty for the orientation. A Lie group $\mathcal{G}$ is a smooth manifold $\mathcal{M}$, which locally resembles a linear space. This implies that there exists a unique Euclidean tangent space $\mathcal{T}_\mathbf{Y}$ at each point $\mathbf{Y}$ on the manifold. All these tangent spaces have the same structure and therefore, can be transformed into each other. To model probability distributions on a Lie group, one can simply define $\mathbf{Y}$ as a perturbation with $\boldsymbol{\tau} \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathcal{M}$ around the mean $\bar{\mathbf{Y}}$ locally in its tangent space. Thus, $\mathbf{Y}$ and its covariance matrix $\boldsymbol{\Sigma}_\mathbf{Y}$ can be expressed in terms of $\boldsymbol{\tau}$

$$\begin{aligned}
\mathbf{Y} &= \bar{\mathbf{Y}} \oplus \boldsymbol{\tau} \;\; &&= \bar{\mathbf{Y}} \circ \mathrm{Exp}(\boldsymbol{\tau}) \\
\boldsymbol{\tau} &= \mathbf{Y} \ominus \bar{\mathbf{Y}} \;\; &&= \mathrm{Log}(\bar{\mathbf{Y}}^{-1} \circ \mathbf{Y}) \\
\boldsymbol{\Sigma}_\mathbf{Y} &= \boldsymbol{\Sigma}[\boldsymbol{\tau}\boldsymbol{\tau}^T] \triangleq \mathbb{E}\left[(\mathbf{Y} \ominus \bar{\mathbf{Y}})(\mathbf{Y} \ominus \bar{\mathbf{Y}})^T\right] \in \mathbb{R}^{m\times m},
\end{aligned}$$

where $\mathrm{Exp}(\boldsymbol{\tau})$ is the retraction of $\boldsymbol{\tau}$ onto $\mathcal{M}$ and Log is the inverse operation that transfers an element of $\mathcal{M}$ to its tangent space. Therefore, a Gaussian distributed variable on Lie groups can be naturally expressed in the tangent space as $\mathbf{Y} \sim N(\bar{\mathbf{Y}}, \boldsymbol{\Sigma}_\mathbf{Y})$. With this in mind, the correspondence likelihood for the orientation $\mathbf{R}$ becomes

$$p(\mathbf{R}|C) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_\mathbf{R})}} \exp\left(-\frac{1}{2}(\mathbf{R} \ominus \bar{\mathbf{R}})^T\boldsymbol{\Sigma}_\mathbf{R}^{-1}(\mathbf{R} \ominus \bar{\mathbf{R}})\right),$$

where $\boldsymbol{\Sigma}_\mathbf{R} \in \mathbb{R}^{3\times3}$ is the covariance matrix, and $\bar{\mathbf{R}} \in \mathrm{SO}(3)$ is the mean orientation of a hypothesis. Note that the mean values are taken from filtered action hypotheses. Furthermore, we assume that the orientation $\mathbf{R}$ and translation $\mathbf{t}$ are conditionally independent, i. e., $p(\mathbf{R},\mathbf{t}|C) = p(\mathbf{R}|C) \cdot p(\mathbf{t}|C)$. Now, the Bayes' rule can be used to estimate the likelihood $p(C|\mathbf{R},\mathbf{t})$ that the observed action $\mathbf{A}$ at position $\mathbf{t}$ and orientation $\mathbf{R}$ corresponds to the hypothesis $\bar{\mathbf{A}}$

$$p(C|\mathbf{R},\mathbf{t}) = \frac{p(C)\cdot p(\mathbf{R}|C)\cdot p(\mathbf{t}|C)}{p(\mathbf{R})\cdot p(\mathbf{t})} \propto p(\mathbf{R}|C)\cdot p(\mathbf{t}|C).$$

To avoid checking every single hypothesis for correspondence with all observations in a given scene, we run a search with a k-dimensional tree. For every hypothesis, a search radius $r = 3\cdot\sigma_m = \max \mathrm{diag}(\boldsymbol{\Sigma}_\mathbf{t})$ is used. This is warranted, as the position of a hypothesis follows a multivariate normal distribution with independent components and the confidence region defined by the three times scaled *Standard Deviational Hyper-Ellipsoid* is enclosed by the sphere with $r = 3\cdot\sigma_m$. Therefore the probability of finding a sample (i. e., a corresponding observation) inside this sphere is larger than $\sim 97\%$ [27].

2) Bayesian Filtering of Action Hypotheses: Once a corresponding observation $\mathbf{A}_t$ of the filtered action hypothesis $\bar{\mathbf{A}}_{t-1}$ has been identified, the next step is to update the estimated state of the filtered action using that observation. Therefore, an update to the *existence certainties* $p_E^{a_i}$, as well as the pose $\bar{\mathbf{X}}$ of $\bar{\mathbf{A}}$ has to be performed.

The *existence certainty* can be determined using the previously calculated correspondence certainty $p = p(C|\mathbf{R},\mathbf{t})$. For that we use a *Continuous Density Hidden Markov Model* (CDHMM) with 2 states [28]. The hidden states are S_1 (action hypothesis exists) and S_2 (action hypothesis does not exist). As a hypothesis is initially created when an observation indicates that it does exist, we assume that the CDHMM is in S_1 upon creation and that the state can only go from existing (S_1) to not existing (S_2), i. e., a hypothesis can only vanish. Therefore, we can define the state $\lambda = (\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ of the CDHMM with $\boldsymbol{\pi} = \begin{pmatrix}1 & 0\end{pmatrix}$ and

$$\mathbf{A} = \begin{pmatrix} a_{11} & 1-a_{11} \\ 0 & 1 \end{pmatrix},\; \mathbf{B} = \begin{pmatrix} p & 1-p \\ 1-b_{22} & b_{22} \end{pmatrix}.$$

The *forward-backward* algorithm is then used to calculate the probability $p_E^{a_i}$ of being in state S_1 at time t for the affordance a_i.

Since the orientation $\mathbf{R}$ cannot be easily modeled in Euclidean space, a simple Kalman filter is not adequate for the estimation of the pose $\bar{\mathbf{X}}$ from multiple observations $\mathbf{X}$. Recently, multiple works have investigated the use of Lie groups for robotic applications and state estimation [5], [16], [20], [21], as they offer a natural, smooth representation of poses.

In [5], Gaussian distributions on manifolds are used to implement a UKF for generic Lie groups using retractions onto the tangent space, where the standard algorithms for a UKF can be used to update and propagate the state. We use the open-source implementation of a UKF on manifolds (*UKF-M*) [29] for the spatio-temporal fusion of the mean pose $\bar{\mathbf{X}}$ of the filtered action $\bar{\mathbf{A}}$ and its covariance matrix $\boldsymbol{\Sigma}_\mathbf{X}$ from multiple action observations $\mathbf{A}$.

Together with the previously defined fusion of the existence certainty for affordances, a complete probabilistic state estimation using Bayesian statistics for actions is possible.

IV. Experiments

We evaluated the approach regarding the spatio-temporal fusion of action candidates for the *graspability* affordance in a series of real-world grasping experiments on the humanoid robot ARMAR-6 in two cluttered setups: (1) Box emptying

setup and (2) table clearing setup. In both setups, unknown objects are randomly placed and should be grasped and manipulated to empty the box or clear the table. A video describing the approach and experiments can be found under https://youtu.be/lXxWtTIySB0

A. Setup

Top-grasp candidates in both scenarios were generated using three methods: (A) the *geometry-based action extraction* (*GAE*), using surface patch-based action observations (utilizing the methods described in Section III-A and Section III-B) without fusing them, (B) the *probabilistic action extraction and fusion* (*PAEF*) described in Section III-A through Section III-C, and (C) the approach based on *object-oriented bounding boxes* (*OOBB*) combined with a region growing segmentation described in our previous work [30].

The parameters used for the *PAEF* method were chosen empirically. For the supervoxel clustering, the parameters $\alpha = \beta = \gamma = \delta = 5$ were used and their respective distances were normalized, so that all contributions were on the same scale. For the Kalman filter, the initial position covariance was chosen to be $1\,\mathrm{cm}$ and the initial orientation covariance to be $0.1\,\mathrm{rad}$. The parameters of the HMM are chosen as $a_{11} = 0.9$ and $b_{22} = 0.5$. After generation of the grasping action candidates, each candidate was checked for a feasible solution of the inverse kinematics and, depending on the orientation of the grasp candidate, a suitable placement of the mobile robot base, as well as one of the robot's hands to execute the grasp with, are chosen. From all valid grasp candidates, the highest one was selected, executed, and stored for reference. Prospective candidates are preferred if they do not lie inside a small area around each previous candidate to increase the variability of the executed grasps and to prevent that a single grasp is executed multiple times in succession. The reaching motions for the execution were generated using *Via-point Movement Primitives* [31] which allow adaptation to new goals, i. e., grasping poses.

B. Box Emptying Experiments

The first experiments are conducted in a setup similar to the one used in our previous work [7]. A varying number of unknown objects used in a decontamination scenario are randomly placed inside a box, 30 grasp attempts were performed for each candidate generation method (*GAE*, *PAEF*, and *OOBB*) and the results were recorded. This was repeated in five setups with a varying number (ranging from 6 to 14 per setup) of randomly placed objects in the box. The objects consist of boxes, cylinders, and other more complex shapes, like bent pipes, a hammer, or a spray bottle. To increase the variability of the investigated scenes, during the 30 grasps the object configuration was changed after 5 grasps attempts by either rearranging or even exchanging the objects in the box by a human operator. In case of a successful grasp, the object was lifted and dropped from about 30cm height before executing the next grasp. Through the robot's interaction with the scene and the frequent rearranging of the objects, we aim at reducing bias regarding the generation of object configurations in the scene in all setups.

Outcome	Description
Grasped	The object does not touch the ground for 5 seconds
Stable Lifted	The object is lifted for 5 seconds but parts of the object still touch the ground
Lifted	the object is visibly lifted for less than 5 seconds
Collision	The object is not lifted because the hand collides with other objects or the environment (e. g., box)
Slipped	The object is not lifted because the hand slipped off the object / was misaligned
Missed	The grasp is generated incorrectly, no object is close enough to be grasped or no executable grasp is found after 2 minutes

TABLE II: Possible outcomes of grasping attempts.

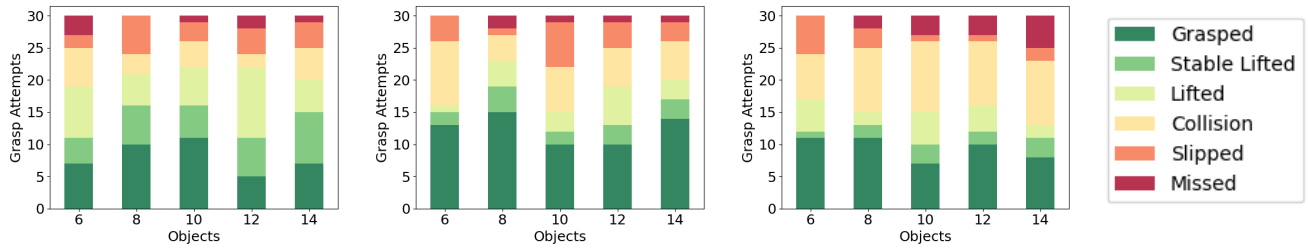
The grasp attempts were categorized based on the result of the execution and failure reasons, as can be seen in Table II. Additionally, for each grasp attempt, the time from candidate generation until candidate selection was measured. Since the *PAEF* and the *OOBB* grasp generation methods rely on previous scene observations to calculate new candidates, the methods were reset after each grasp attempt to provide a worst-case estimate for the time required for the generation and selection of grasp candidates in novel scenes.

C. Table Clearing Experiments

The second experiment was conducted in a kitchen scenario and comprised clearing an $80\,\mathrm{cm} \times 80\,\mathrm{cm}$ table cluttered with 18 objects and stowing them in a box. The objects consisted of 6 boxes, 3 cups, 3 plates, and 6 fruits like pears, lemons, and oranges. If a grasp attempt fails or an object is dropped, the grasping process was manually aborted and restarted by a human operator. The grasp candidate generation, selection, and execution were done in the same manner as in the box emptying scenario. For each grasp generation method, the table was cleared five times and the number of objects in the stowing box, as well as the total time and number of grasp attempts necessary to clear the table, was recorded for each experiment. An experiment ends if all objects were removed from the table or if no executable grasp is found for the duration of 5 minutes.

D. Results and Discussion

1) Box Emptying Experiments: The results of the grasping experiments are shown in Figure 4. If only the cases "grasped" and "stable lifted" are counted as successful grasp candidates, the grasp extraction using only action observations (*GAE*) has an average success rate of 46.0%, while the *OOBB* method has only 38.7%. The additional spatio-temporal fusion of the actions (*PAEF*) increases the success rate to 50.7%. For the successful executions, there is no strong correlation on the number of objects in the box for all methods with a Pearson's correlation coefficient of $\rho_{PAEF} = -0.11$ for *PAEF*, $\rho_{GAE} = 0.18$ for *GAE* and $\rho_{OOBB} = -0.42$ for *OOBB* candidates. On the other hand, only for the *OOBB* candidates,



(a) *GAE*-based Grasping (b) *PAEF*-based Grasping (c) *OOBB*-based Grasping

Fig. 4: Results of the box emptying experiments on ARMAR-6

there is a correlation for missed grasps and the number of objects in the box with $\rho_{GAE} = -0.28$, $\rho_{PAEF} = 0.22$ and $\rho_{OOBB} = 0.96$. This shows, that while all methods are able to generate sufficiently good grasping candidates even in very cluttered scenes, *PAEF* and *GAE* were able to consistently perform well in the most difficult scenes. This confirms our hypothesis that the use of only local surface geometry, independent of the segmentation of the point cloud, has a positive influence on the accuracy of our approach. As the average grasp success rate of the *PAEF* method increases by more than 4% in comparison to the *GAE* method, it is evident that the spatio-temporal fusion of actions has a positive impact on the robustness of our approach. On the other hand, *OOBB* has fewer failures than *PAEF* due to slipping of the object, which happens mostly when there is a slight offset in the pose from where an "ideal" grasp would be located. Therefore, it seems that in cluttered scenes action observations are fused that do not belong to one object or action hypothesis. This is not the case for *OOBB*, as the candidates have access to global context in the form of the bounding boxes that are supposed to span an entire object. Most of the failed grasp attempts were caused by the collision of the hand with other objects during the reaching or hand closing phase. This is especially visible for the *OOBB* candidates, which can be explained by fewer generated candidates. In several situations, only one reachable candidate was generated by *OOBB*, which resulted in the same candidate being executed multiple times. If the execution does not lead to significant change of the scene, the grasp will fail again due to the same reason. In some object configurations in the box, the *OOBB* method was not able to find a grasp at all and required manual changes to the scene to recover.

Due to an increasing number of action observations being generated as the number of objects in the box increases, the *PAEF* method generally requires more time for grasp selection than the *OOBB* method, as can be seen in Figure 5. This can be explained by the longer time needed for the search for action correspondences, as a larger number of hypotheses has to be checked and subsequently fused. As the *OOBB* method generates fewer candidates, only dependent on the number of point cloud clusters segmented in a scene, with an averaging over close candidates, the time needed to generate candidates is almost constant over all degrees of clutter. The *GAE* method requires approximately the same time for all setups, as the number of surface patches to be processed is independent of

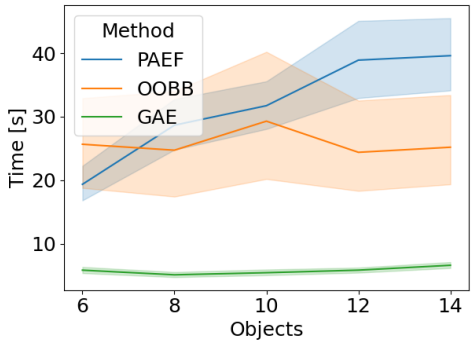


Fig. 5: Comparison of the combined extraction, filtering, and selection times for the *GAE*, *PAEF* and *OOBB* methods

the number of objects in the scene. It is faster than the other methods, as no averaging or filtering is performed.

2) Table Clearing Experiments: The results of the table clearing experiment can be seen in Table III. As in the box emptying scenario, the success rates of the *GAE* and *PAEF* methods exceed those of the *OOBB* method. While the total amount of objects stowed away for the *PAEF* method is only slightly larger than for the *GAE* method, the number of necessary grasps to clear the table is by far higher for the *GAE* method. This is in accordance with the results of the box emptying scenario and stems from the higher robustness of the temporally fused grasp candidates. Additionally, the higher accuracy of the *PAEF* method partially compensates for the higher extraction times, as the total time to clear the table was only slightly larger than for *GAE* method. The *OOBB* method

	GAE	***PAEF***	***OOBB***
Stowed Boxes	4.6 ± 1.7	5.6 ± 0.5	4.2 ± 0.8
Stowed Plates	2.8 ± 0.4	2.0 ± 0.7	1.6 ± 1.1
Stowed Cups	2.0 ± 1.0	2.8 ± 0.4	2.2 ± 0.8
Stowed Fruit	2.4 ± 1.1	1.8 ± 0.8	1.8 ± 1.3
Total Stowed	11.8 ± 1.5	12.2 ± 0.8	9.8 ± 1.1
Remaining Objects	0.8 ± 0.8	1.6 ± 1.1	5.2 ± 2.4
Grasp Attempts	37.0 ± 2.2	30.2 ± 4.2	29.4 ± 6.1
Total Time [min]	33:11 $\pm$ 2:31	34:39 $\pm$ 5:39	25:24 $\pm$ 2:50

TABLE III: Results of the table clearing experiments.

only has a small number of grasp attempts and a low clearing time, as an average of more than 5 objects remained on the table when the experiment had to be terminated due to the time constraint.

V. Conclusion

In this work, we presented an approach for the extraction of scene affordances based only on the local surface geometry of a point cloud and the subsequent probabilistic, spatio-temporal fusion of the corresponding grasping and manipulation action candidates. To this end, we defined a geometry-aware shared state for all affordances in the form of the *Local Curvature Frame* and used methods from *Bayesian Recursive State Estimation* to fuse the pose of this frame over multiple distinct observations. Based on the averaged *Local Curvature Frame*, concrete action candidates can be synthesized in a global coordinate system. The approach was tested in multiple real-world grasping scenarios on the humanoid robot ARMAR-6 and compared to the grasp extraction based on the method presented in [30] in more than 900 grasp executions. Our approach performed consistently better than a grasp candidate extraction based on bounding boxes of the unknown objects over all degrees of clutter in a scene and resulted in an increase of over 10% in grasp success rate. We also proved that the success rate is largely independent of the number of objects in the scene, and therefore, the approach is able to handle difficult and cluttered scenes. We postulate that this is because our approach is independent of the scene segmentation.

In the future, we plan to use global information for the correspondence search, as there were indications of wrong correspondences being fused during the experiments in scenes with a higher number of objects. Furthermore, we want to extend our approach to other affordances, as we only consider grasping affordances in this work, and test the entire framework in more challenging scenarios.

References

[1] J. J. Gibson, "The theory of affordances," in *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979, ch. 8, pp. 119–137.

[2] P. Kaiser, E. E. Aksoy, M. Grotz, and T. Asfour, "Towards a hierarchy of loco-manipulation affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2016-Novem, 2016, pp. 2839–2846.

[3] P. Kaiser and T. Asfour, "Autonomous Detection and Experimental Validation of Affordances," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1949–1956, 2018.

[4] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, Mass.: MIT Press, 2005.

[5] M. Brossard, S. Bonnabel, and J.-P. Condomines, "Unscented Kalman filtering on Lie groups," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2017-Septe, sep 2017, pp. 2485–2491.

[6] T. Asfour, M. Wächter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "ARMAR-6: A High-Performance Humanoid for Human-Robot Collaboration in Real World Scenarios," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[7] C. Pohl, K. Hitzler, R. Grimm, A. Zea, U. D. Hanebeck, and T. Asfour, "Affordance-Based Grasping and Manipulation in Real World Applications," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, oct 2020, pp. 9569–9576.

[8] K. M. Varadarajan and M. Vincze, "Affordance based Part Recognition for Grasping and Manipulation," *ICRA Workshop on Autonomous Grasping*, 2011.

[9] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation," *International Symposium on Robotics Research*, 2019.

[10] P. Ardon, E. Pairet, R. P. Petrick, S. Ramamoorthy, and K. S. Lohan, "Learning Grasp Affordance Reasoning through Semantic Relations," *IEEE Robotics and Automation Letters*, vol. 4, pp. 4571–4578, 2019.

[11] B. Moldovan, P. Moreno, D. Nitti, J. Santos-Victor, and L. De Raedt, "Relational affordances for multiple-object manipulation," *Autonomous Robots*, vol. 42, no. 1, pp. 19–44, 2018.

[12] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae, 30th Edition*. CRC Press, 1995.

[13] T. Tsuji, S. Uto, K. Harada, R. Kurazume, T. Hasegawa, and K. Morooka, "Grasp Planning for Constricted Parts of Objects Approximated with Quadric Surfaces," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 2447–2453.

[14] D. Kanoulas, C. Zhou, A. Nguyen, G. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Vision-based foothold contact reasoning using curved surface patches," in *IEEE-RAS International Conference on Humanoid Robots*, 2017, pp. 121–128.

[15] T. D. Barfoot, *State Estimation for Robotics*. Cambridge: Cambridge University Press, 2017.

[16] J. Solà, J. Deray, and D. Atchuthan, "A micro Lie theory for state estimation in robotics," *arXiv:1812.01537*, 2018.

[17] S. Bultmann, K. Li, and U. D. Hanebeck, "Stereo Visual SLAM Based on Unscented Dual Quaternion Filtering," in *International Conference on Information Fusion*, 2019.

[18] E. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *IEEE Adaptive Systems for Signal Processing, Communications, and Control Symposium*, 2000, pp. 153–158.

[19] G. Loianno, M. Watterson, and V. Kumar, "Visual inertial odometry for quadrotors on SE(3)," in *IEEE International Conference on Robotics and Automation*, no. 3, 2016, pp. 1544–1551.

[20] A. M. Sjøberg and O. Egeland, "Lie Algebraic Unscented Kalman Filter for Pose Estimation," *IEEE Transactions on Automatic Control*, 2021.

[21] T. Lee, "Bayesian Attitude Estimation with the Matrix Fisher Distribution on SO(3)," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3377–3392, 2018.

[22] N. M. Patrikalakis and T. Maekawa, *Shape Interrogation for Computer Aided Design and Manufacturing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[23] A. Spek, W. H. Li, and T. Drummond, "A Fast Method For Computing Principal Curvatures From Range Images," in *Australasian Conference on Robotics and Automation (ACRA)*, 2015, pp. 33–41.

[24] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2027–2034.

[25] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation*, 2011.

[26] J. Bohg and D. Kragic, "Learning grasping points with shape context," *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 362–377, 2010.

[27] B. Wang, W. Shi, and Z. Miao, "Confidence Analysis of Standard Deviational Ellipse and Its Extension into Higher Dimensional Euclidean Space," *PLOS ONE*, vol. 10, no. 3, 2015.

[28] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Readings in Speech Recognition*. Elsevier, 1990, pp. 267–296.

[29] M. Brossard, A. Barrau, and S. Bonnabel, "A Code for Unscented Kalman Filtering on Manifolds (UKF-M)," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 5701–5708.

[30] R. Grimm, M. Grotz, S. Ottenhaus, and T. Asfour, "Vision-Based Robotic Pushing and Grasping for Stone Sample Collection under Computing Resource Constraints," in *IEEE International Conference on Robotics and Automation*, 2021.

[31] Y. Zhou, J. Gao, and T. Asfour, "Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 4301–4308.