The importance of structure

Carl Henrik Ek and Danica Kragic

Abstract Many tasks in robotics and computer vision are concerned with inferring a continuous or discrete state variable from observations and measurements from the environment. Due to the high-dimensional nature of the input data the inference is often cast as a two stage process: first a low-dimensional feature representation is extracted on which secondly a learning algorithm is applied. Due to the significant progress that have been achieved within the field of machine learning over the last decade focus have placed at the second stage of the inference process, improving the process by exploiting more advanced learning techniques applied to the same (or more of the same) data. We believe that for many scenarios significant strides in performance could be achieved by focusing on representation rather than aiming to alleviate inconclusive and/or redundant information by exploiting more advanced inference methods. This stems from the notion that; given the "correct" representation the inference problem becomes easier to solve. In this paper we argue that one important mode of information for many application scenarios is not the actual variation in the data but the rather the higher order statistics as the *structure* of variations. We will exemplify this through a set of applications and show different ways of representing the structure of data.

1 Introduction

A central question to solve when designing an artificial system is how to make it aware and capable of interaction with the environment. The level of usefulness of a robot is considered through its capability of reacting to and adjusting its behavior to changes in the environment. Todays robots, equipped with different sensors such as cameras, microphones and depth sensors acquire information from the environment at very high precision and rate. Through this rapid development it is now possi-

Carl Henrik Ek
Royal Institute of Technology, Sweden, e-mail: chek@csc.kth.se

Danica Kragic
Royal Institute of Technology, Sweden e-mail: dani@csc.kth.se

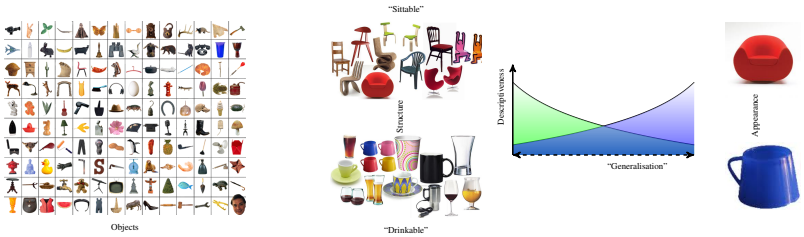

Fig. 1 *The above figure tries to highlight the notion of the importance of structure that we try to convey in this paper. The example above shows a large data-base of objects to the far left. Of these we want find a representation in order to classify objects at a certain resolution. If the representation naturally generalizes, i.e. it does not reflect within class variance but only between this task is easy to solve. In this paper we argue that for a coarse scale task such as separating "sittable" from "drinkable" objects the discriminating variance is represented by the global structure. While for a high resolution task such as separating the "red felt comfy chair" or the "blue plastic mug" the discriminating information is contained in the appearance cues. We believe that in robotics we are generally interested in the first type of these two task why therefore find representations of global structures is important.*

ble to design artificial systems whos sensory systems are more capable than those of the human. However, despite getting more and more detailed observations of the environment, the progress in what we are able to infer through reasoning from this data have not seen the same rapid development. Our central argument in this paper is, *Given the "right" information about a domain inferring the correct answer becomes an easier problem*. The development of sensory systems have rather than focusing on providing the *"right"* information been aimed at simply acquiring *more* information. The justification for this has been the development of more and more advanced machine learning algorithms capable of dealing with larger amounts of data of more complicated distributions. However, the fact still remains that the progress in terms inference have not followed that of the sensory systems.

One of the strengths of human inference is its capability of being selective with the information it uses to reason [1]. During our development we construct strong (conditional) priors which helps us filter the enormous amount of information that our sensory systems acquires to only use a small subset of the data which is relevant for the task, as indicated by the concept of intentional blindness shown in [2]. Rather the opposite approach seems to be dominant when building artificial systems where we try to extract and model more and more of the variations in the sensory data and exploit more advanced learning algorithms for inference from a very complicated input domain. A describing example is object cathegorisation in computer vision where the dominant approach is to use local image descriptor such as SIFT [3] to model the sensory data. Clearly the information extracted by such features contains significant amounts of variance which is not relevant for the task which means that in order to be able to generalize within categories the inference algorithm needs to learn to ignore data and focus on the discriminating information. In many situations the discriminating information represents only a small part of the variance in

the extracted representation which often means a significant challenge in terms of modeling and inference.

In this paper we argue that rather than focusing on building models capable of representing a larger amounts of the variance in the sensory, we should aim to carefully consider what information that is actually relevant. We argue for representations that focus on the structure of variations rather than accurate descriptions of the local variations in the data. Our motivation stems from the notion that the biggest challenge when it comes to inference is not discrimination per say but rather its complementary notion that of generalization. I.e. the key problem is not to extract variance that separates certain classes but rather avoid extracting variance that corresponds to within class variations. As an example, having observed a specific instance of a mug we can reasonably reliably detect that mug again, the big challenge is to create a system which is capable of generalizing over different mugs separating them from other objects.

We argue that the important questions are concerned with generalization on a level where the global structure is the dominant discriminating factor and not the local variations see Fig. 1. To that end we will describe a set of different scenarios where structural representations and models are of key importance. Through these examples we will show different approaches for exploiting global structure. However, we would like to point out that the purpose of this paper is not to provide a solution to a specific problem but rather exemplify our notion through a range of applications where structure is important to stimulate further discussion on the subject.

2 Structure and Generalization

There are three central concepts in this paper; those of *generalization*, *discrimination* and that of *structure*. To explain what we mean by these we use the example of object modeling as this provides a intuitive example of the concepts that we address in this paper. Object modeling is a necessary prerequisite for equipping a robot with the ability of detection, identification and manipulation. Dependent on the task, we wish to acquire a representation that generalize over specific classes and is able to discriminate between others. Formally this means that we wish to model the between class variance but not the within. Thus, the two concepts generalization and discrimination are complementary, but from a traditional representation point of view the biggest challenge is not to retain (the discriminative part) but rather to remove (the generalizing part) information. An example of this is representing object from visual data for the task of categorization. The main challenge is not to find a representation that separates for example mugs from glasses, as they look different the information is contained in the observations, but rather to remove the information that separates different mugs and different glasses from each other.

Any statistical method relies on the presumption that we can acquire enough samples of a space that can describe it well. Images are high-dimensional meaning that it is not possible to acquire such a data-set easily. The traditional approach has been

to extract a low-dimensional feature representation assuming that we can acquire samples that describe the feature space. The most obvious approach is to extract this information from a local patch in the image as clearly this will per definition contain less variations. The central question is then: What level contains the desirable generalization and discrimination characteristics for a specific task? Clearly, on the most local level, being the colour of a pixel, we can model the information robustly and the assumption of sampling the feature space well is going to be forefilled by observing a single image. However, we also know that statistics of such local features will not contain discriminating information for other than the most simple task while it will generalize over a large range of different images. Here is an important notion: the more local a feature, the less discriminative it becomes. Thus, there is a trade-off here that needs to be considered, local enough to be robust and well sampled and global enough to be descriptive, see Fig 2.

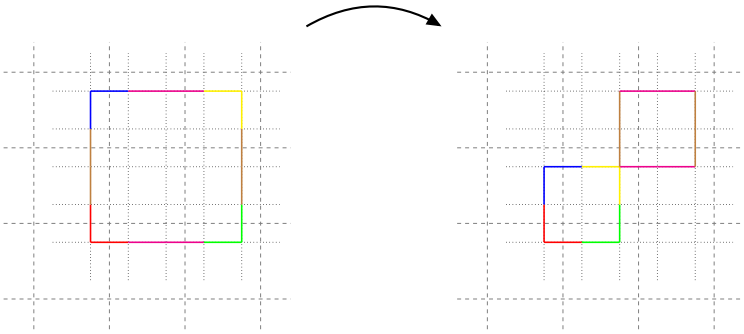

Fig. 2 *The above figure shows two different objects with two different scales of local representation, dotted (fine) and dashed (coarse). First order statistics from the fine resolution will not be able to discriminate the two objects while at the coarser scale they will be different. However, using a coarser scale implies that each cell has a higher dimensionality requiring more samples in order to represent the space well.*

The traditional approach have been to try and use more and more descriptive local features by acquiring large (and growing!) training data sets and then exploit a supervised machine learning technique capable of learning a secondary representation, often through the use of kernel machines or metric learning, that achieves the desired balance between generalization and discrimination.

We argue that the there is a different paradigm where we could use less informative local descriptors while still being able to discriminate. That is to aim to create strong models of the *structure* between the local features and not stop at first order statistics such as the so popular *Bag-or-words* techniques. However, how to encode structure is a non-trivial problem that we believe needs to be addressed with much more focus. We do not think that there is one single approach for representing structure but rather a large range of different tools and approaches. In the reminder of this paper we will show different applications and different intuitions and tools that are useful and provide insights into how to deal with different tasks by including a

structural element. Our goal with this paper is rather to raise questions than provide specific solutions.

3 Temporal Structure

Many tasks in robotics deal with dynamical scenes where the relevant information is contained in the order of events. A goal of robotics is learning by demonstration [4] where the task is for a robot to extract the relevant notion of a task by observing a demonstrator. Various subproblems have been studied related to task planning and sequencing, detection of motion primitives, developing models for structured collections of actions [5]. The underlying question has been how to acquire a representation that in a sufficient manner generalizes the objective(s) of the task. Take for example the task of clearing a table. Here the appearance of both the objects and the table are irrelevant. Rather the important information that generalizes the task lies in the structure of the events not the actual events themselves. I.e. the task remains the same if the cuttlery are cleared before the plates or vice-versa. In this section we describe different applications where we, through some model of temporal structure, manage to simplify an otherwise complicated inference task.

3.1 Interaction

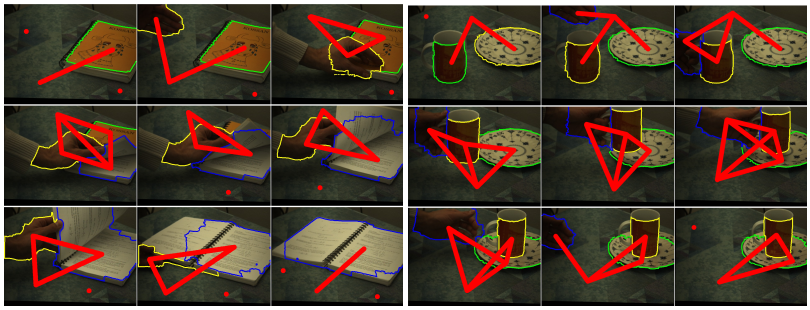

Fig. 3 *The* **left** *example shows an instance of the* **Opening Book** *action while the* **right** *shows the* **Moving Object**. *In each of the images the result of the segmentation and its corresponding graph have been overlaid . Only the spatial relations between the segments are extracted and no identification of the objects is performed.*

Recently, [6] suggested a method for action classification by representing the temporal structure of the interactions that takes place in the scene. Using visual measurements from a camera the approach first segments the objects in the scene for each frame in a video sequence. The temporal structure is encoded by a graph representing each frame, every object being a node and connected component shar-

ing an edge, see Fig 3. This process removes all information associated with the appearance and identity of the objects leaving only the interaction. The final processing step is to remove the duration of the interactions and only retain the sequence of topologically different graphs. The intuition behind the representation is that for discriminating between actions the temporal structure of the interactions of objects independent of their identity contains sufficient information. This is significantly different to the more traditional approach for modeling actions such as [7, 8, 9] which extracts a representation that retains a significant amount of the variance related to appearance. This means that we have to learn the invariance related to appearance from data. This requires significantly larger amounts of training data and puts additional challenges on the learning machinery that needs to explain away this non-relevant variance and extract the important variance from the feature. In order to represent each frame the authors in [6] defines a specific semantic extracted from the the node connectivity in the graphs and the alterations under this semantic over time is represented as a matrix. A simple distance measure is then defined to compare two different matrices which given a training data-set allows for action classification.

One of the major drawbacks of the approach suggested in [6] is that it is very sensitive to noise as it assumes that each node in the graph represents a single object. In order to circumvent this problem, we have developed a general framework for encoding the structure of variation in a semantic chain using a robust machinery derived from work in text representation [10]. We are motivated by the approach presented in [11] where a feature space representation of a string is presented. By deriving a vector space representation of a string independent of its length strings can be compared by standardized tools from statistical learning. The parametrisation is sensitive to both the order and the existence of letters in the string and does therefore encode both the structure and the appearance of the string. Being infeasible to compute for most typically sized data-sets the feature space is represented implicitly through the use of a kernel function [12]. More formally the feature space we use is spanned by all possible permutations of all lengths of the letters in the semantic alphabet, with an inner product defined as a function of the matching part of the overlap between two strings, see Fig 4. Clearly the spaces is infinite dimensional but as any string of a shorter length compared to the basis are orthogonal the maximum dimensionality is bounded. Similarly to the original string kernel [11] an efficient recursive computation of the inner product can be formulated representing the feature space implicitly using by a kernel.

The above example completely removes all variance associated with appearance from the observations and only retains information about structure. For the task of discriminating between the different actions defined in [6] this contains sufficient information. However, it is easy to think of scenarios where this information is not sufficient for performing the task. However, the kernel based framework can easily be adapted to encode structure where the appearance is also retained as this is simply about defining a semantic that also encodes the appearance. As an example of such we will describe an approach for representing object categories that retain both the

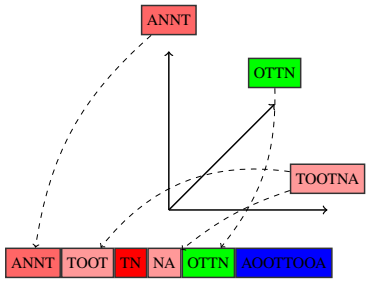

Fig. 4 *For a the specific semantic alphabet, here defining the four different interaction relationships between objects:* $\{A,N,T,O\}$*, we above show a subspace of the feature space representing the sequence. The sequence* **ANNT** *(red) and* **OTTN** *(green) exists in order in the string and will therefore project parallel to the corresponding basis while the* **TOOTNA** *does not which will induce a non-zero angle between the string and the basis. This means that the representation will be sensitive to gaps in the string making it robust to noise.*

Fig. 5 **Left** *The bar plot above shows the classification rate associated with increasing noise to the right. The* **green** *bars identifies our kernel approach while the* **red** *indicates the performance of the original method.* **Right** *Confusion matrices for increasing noise. The classes are ordered as* **Moving Object**, **Making Sandwich**, **Opening Book** *and* **Filling Liquid**. *The red matrices show the results for the original approach while the results of our method is shown in green. With increasing amount of noise the original measure is unable to disambiguate between the different actions classifying every action as belonging to opening book. For the same data the kernel approach is able to differentiate between the classes and the performance is reduced much more gracefully.*

appearance and the structure of the object. An idea for the future is the integration of this approach with the probabilistic models for action encoding presented in [13].

3.2 Object Detection

A robot should be able to interact with it surroundings by applying actions to objects. Thus, a very important task is to identify and extract objects from sensory data. The visual domain contains a rich description of the environment and by segmenting objects from the background detailed models of individual objects can be built. Image segmentation is concerned with clustering "similar" pixels into segments and has attracted considerable interest in computer vision. There are many different approaches and assumptions used to define similarity between pixels. Be-

cause of computational limitations, but also due to the challenge of formulating general appearance models, the focus has been on local statistics such as colour distributions and gradients [14, 15]. This has meant that for all but the simplest objects it is quite unlikely that the clusters retained by an image segmentation approach will corresponds to actual objects in the scene.

The work in image segmentation shows the non-trivial nature of formulating consistent appearance cues based on local statistics that corresponds to objects in the image. This has meant that most successful approaches are interactive, requiring a human to refine and rectify the result produced in an iterative manner [15]. In an autonomous system we cannot rely on interaction to leverage and include human object priors for segmentation but rather need to create a self-contained system.

In [16] we presented an active system for object segmentation which exploits both traditional appearance based assumptions in collaboration with temporal cues in an active iterative manner. Image segmentation techniques are good at grouping pixels into consistent regions. This often mean that for all but the simplest objects this will result in an *over* segmentation where each object is divided into several different segments. Acknowledging the fact that it is a non-trivial task to create appearance models that encapsulates the long range pixel interactions that generalizes over objects we turn our attention to a different domain. In many applications we can assume that the objects of interest in the scene are rigid. Further, each local element or point on such an object moves according to simple rules of rigid motion. This means these rules *generalizes* over all points belonging to the same object. To that end we use the initial segmentation from the appearance cues as an hypothesis of the objects in the scene. In correspondence with this the robot introduces motion by interacting with the scene. Modeling the motion we can easily verify if the appearance segmentation is consistent with the ridig motion assumption. In [16] we describe which combines local appearance cues with a method for modeling rigid motion to use them in a complementary fashion. We show results for a common tabel-top scenario where and appearance based method used on its own would fail, Fig 6.

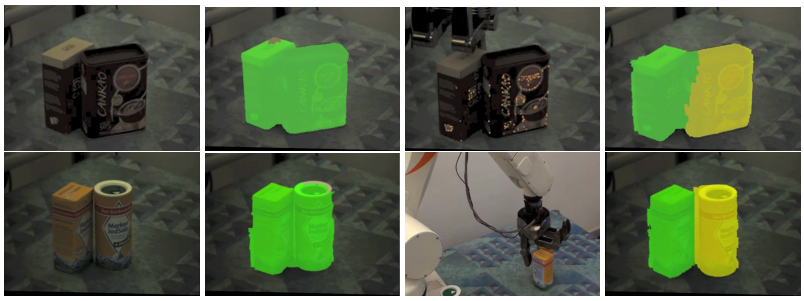

Fig. 6 *The left most column shows two scenarios where two objects have been placed on a table top. Using a traditional appearance based image segmentation approach it is not possible to separate the objects. By introducing motion in to the scene by letting the robot interact with the environment the motion can be modeled and the objects separated in the right most image.*

This approach shows how by exploiting a simple assumption we can actively introduce a variance corresponding to the level of generalization we are interested in such a manner that it can easily be extracted from the environment.

4 Spatial Structure

In previous section, we described applications and tasks exemplifying the importance of temporal structure. In this section we discuss structure on a different level namely the structure on a spatial level.

Similarly to the temporal case we argue that the interesting generalization for many tasks are represented by structural information. On example is our use of language, where we would use an structural adjective such as *striped* to discriminate on a coarse level while for identifying specific objects we would add local appearance descriptions such as *red and white*. The currently dominating approach is to use a local representation of each instance and hope that the inference procedure is capable of extracting the information that generalizes between the classes by observing enough examples. As we have previously stated this is a very challenging task from a learning perspective, as quite likely only a small portion, if any, of the variance in the local descriptor will contain generalizing information.

In this section we describe two different task where the generalizing information is contained in the spatial structure of the local appearance and not the local appearance itself.

4.1 Object Representation

Being able to discriminate between objects both on category and instance level is of key importance for a wide range of task in robotics. This requires an object representation that is capable of generalizing over the desired task dependent domain. In computer vision object cathegorisation has attracted a significant interest. Especially in recent years with the collection of public datasets and high profile competitions such as the Pascal VOC challenge [17]. A large range of different techniques have been applied to the problem where the dominating approach is to aim to extract discriminating information from local image descriptors by relying on the capabilities of different machine learning approaches.

Compared to computer vision researchers roboticists enjoy the luxury of being able to apply several different types of sensory streams in addition to cameras for extracting information of the environment. Recently with the introduction of affordable depth sensors has allowed us to consider dense depth information not as a specialsed domain but rather something that can be assumed as readily available. In [18, 19] a robust $3D$ feature is presented which represents each local patch of an object as belonging to a specific geometric class. In Figure 7 the feature is shown extracted from a set of typical household items. Clearly, only describing the geometrical local structure on the object is not likely to provide discriminative infor-

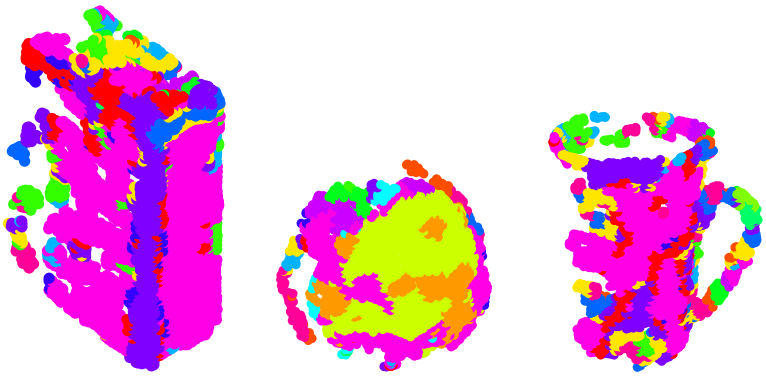

Fig. 7 *Object features representing the local geometrical class encoded by colour shown for three different objects, from the left box, citrus fruit and mug.*

mation between a large range of different object why the global structure needs to be encoded. To that end in [20] the author presents an approach to encode the global structure by encoding the distribution of local patches along rays between each patch.

The results presented are impressive but modeling the distribution of geometrical classes between local patches is not going to retain the full structure of the object and in order to be able to scale in terms of the level of generalization we believe that a stronger representation is needed. In specific we do not think that rays are a good way of encoding the structure of a surface. The objective is to find a representative global statistics that encodes the structure of the object. What we mean in formal terms is that: an object is a two dimensional surface embedded in a three dimensional space which encapsulate a non-empty volume. This implies that given a point on the object one can travel to any other point belonging to the object by traversing this enclosing surface. It is the shape of this surface is what we wish to represent. In this notion of a surface lies our objection towards the use of rays. The surface is a two dimensional object meaning that relating two points to each other requires two degrees of freedom. The position along a ray does not respect the shape of the surface but is rather a construction to create a simple measure of sampling the three dimensional volume along a single parameter. By defining a path respecting the surface of the object, such as the use of an approximate geodesic [21], this defines a distance between each point that reflects the shape of the surface of the object. This distance induces an ordering of each local patch and by representing this ordering rather than the non-surface respecting ordering induces by a ray we believe a more descriptive representation can be found.

Given that we can sample statistics of the object along paths that reflect the true global structure of the object the question remains what type of statistics should be encoded. The obvious approach would be to encode only first order statistics such as in [20] as it can be done in a robust manner and is less sensitive to difference is

sampling resolution. However, we believe that the important information is in the ordering of the local patches not simply the distribution. To that end we wish to take a similar approach as in [10] and exploit robust and principled kernel approaches representation and inference. In specific, where the semantic in [6] does not reflects the local appearance we wish to exchange the semantical alphabet to use the local representation presented in [18]. Rather than modeling the interaction between segments in time we aim to model the interaction spatial, where the time domain is replaces with a distance measure along the object. We believe that this approach has the potential of improving object cathegorisation and classification in a similar manner as it improved action classification as shown in [10]. Our intuition why this will lead to improvement is two fold; only modeling the local structure we are likely to need a very detailed descriptor which is likely to be susceptible to noise. By using a less descriptive local feature as [18] we believe this can be avoided. Secondly, the generalization and discrimination will be encoded by using the robust string kernel approach developed in [10] allowing us to exploit principled and robust inference algorithms for classification.

5 Data Conditional Dependence and Factorization

The previous examples we have discussed have addressed representation of data for a specific problem where we argue that the global structure of the variations in the observations is the key component to model and represent not the actual variations themselves. In this section we will describe a more general case where we do not have a specific task in mind but rather want to acquire a complete model of the data and model its underlying distribution.

In many scenarios of robotics we are given observations of the environment in a factorised form. This can either be that the observations naturally factorises describing separate modalities or through the use of different sensors and or feature representations. Assuming that the observations of the environment $\mathbf{Y}$ factorises into k separate terms $[\mathbf{Y}_1, \ldots, \mathbf{Y}_k]$ this means that from a probabilistic view point the complete model of the environment is represented by the joint distribution, $P(\mathbf{Y}) = P(\mathbf{Y}_1, \ldots, \mathbf{Y}_k)$. However, for many scenarios in robotics the dimensionality of this distribution makes it intractable to learn. In order to proceed one can exploit conditional independence in the observations imposing a structure on the joint distribution such as,

$$P(\mathbf{Y}) = \prod_{i=1}^{k} P(\mathbf{Y}_k | \pi_k), \tag{1}$$

where π_k corresponds to the subspace of $\mathbf{Y}$ that induces a dependency on $\mathbf{Y}_k$ thereby imposing a structure on the observation.

Extracting dependency structures in data is a very hard problem with the number of possible structures growing super-exponentially with the number of variables or nodes. Recently significant strides have been made towards being able to treat structure learning in a principled manner through the development of structural priors such as the Chinese Restaurant Process [22, 23, 24] and Indian Buffet Process

[25, 26]. However, the use of such priors introduces significant limitations on the individual factors in the model meaning that they are not applicable in the general scenario. This means that for many problems researchers have to resort to using heuristic or greedy approaches. Of specific success have been the application of such methods when the data is discrete. However, for most robotic applications we deal with continuous data which means that such approaches have in general been beyond us. As a result, for the general case we often have to assume the structure and or the factorization of the data to be known a priori [27].

In recent [28, 29, 30] work we have created a model which encodes the trade-off between loss of precision as introduced by discretisation process and the benefit of learning the structure by exploiting the heuristic approaches developed for such data. The proposed method learns a continuous latent variable model of each observation space represented by a set of discrete key states. It does so by exploiting recent advances in probabilistic dimensionality reduction [31] and by introducing a specific prior who balances the trade-off between discretisation and representation in a principled manner. In Figure 8 a schematic figure of the graphical model proposed in [28] and the learned intermediate representation used for clustering is shown. Application of proposed method has allowed us to learn the conditional

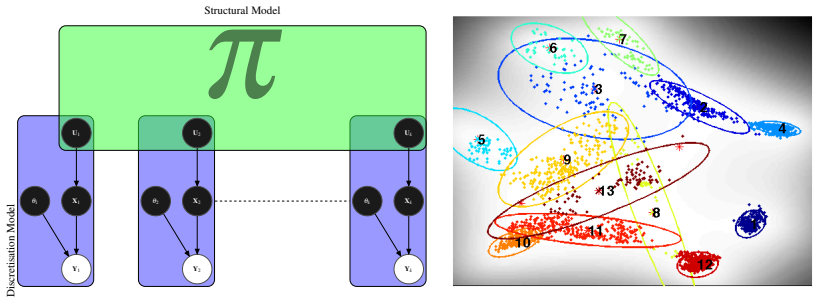

Fig. 8 *The left image shows a schematical graphical model of the structure learning approach. For each continuous observation space* $\mathbf{Y}_i$ *we learn a low dimensional representation* $\mathbf{X}_i$ *with a functional relationship to the observed data parametrised by* θ_i*. Further, the low-dimensional space is represented using a set of discrete locations* $\mathbf{U}_i$*. Given that we have a completely discrete representation in terms of the* $\mathbf{U}_i$ *we can apply traditional heuristic methods for learning the structure* π*. The right image shows and example of the low-dimensional continuous representation and the discretisation colour coded. The separation between the clusters is controlled by a prior modeling the trade-off between discretisation and representation.*

structure from large collections of both discrete and continuous variables within the same model. In Figure 9 the resulting learned structure for modeling a range of different sensor data for a grasping task is shown. This is an example of by enforcing a specific structure on a lower level allows us to learn the more global structure of the data which is often much less trivial to have a notion of. Even though it might not be directly obvious this approach is not particularly different from the previous described methods as: on a lower level we enforce a structure, either in the case here

Name	D	N	Description
task	-	3	Task Identifier
obcl	-	6	Object Class
size	3	8	Object Dimensions
cvex	1	4	Convexity Value $[0,1]$
shcv	3	7	Shape Class Vector (Zernike Similarity)
fcon	20	20	Final Hand Configuration
dir	4	20	Approach Direction (Quaternion)
pos	3	14	Grasp Position
egpc	2	6	Eigengrasp Pre-Configuration
upos	3	11	Unified Spherical Grasp Position
fvol	1	6	Free Volume
gbvl	1	4	Volume of Grasped Boxes
pshcv	3	7	Part Shape Class Vector (Zernike Similarity)
pecce	1	3	Part Eccentricity $[0,1]$
g1bx	1	2	Grasped-1-Box Value $[0,1]$
qeps	1	5	Grasp Stability Measure (eps)
qvol	1	3	Grasp Stability Measure (vol)

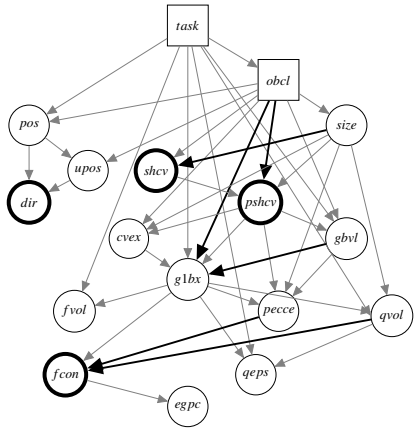

Fig. 9 *Example of a learned factorised representation of* 17 *different observation spaces for a grasping scenario. To the left the different features are shown and to the right the resulting graphical model with the learned structure. The structure is very complicated and it is highly unlikely that we would be able to specify it a priori.*

as a discretisation or in the object category example by on the local feature level extracting specific structures such as edges or face normals, then on a global level we model the structure either as previously in terms of a task or as here in terms of a density model of the data.

6 Topology

Topology is the study of the structure of geometrical spaces and objects. As a branch of mathematics it provides a toolbox for extracting qualitative measurements of geometrical objects. We believe that tools from topology can provide a machinery to encode the global type of structure that we have argued throughout this paper being essential for acquiring a generalizable representation of the environment. However, topology as branch of pure mathematics was not aimed at analyzing uncertain scenarios where we measure the environment through sparse and potentially noisy samples as is often the case in robotics. In [32] the authors argue that by careful consideration of the problem setting, topological tools are applicable to the type of problems where statistical learning have usually been the dominating paradigm. The authors also argue that topological reasoning has the potential to eliviate some of the shortcomings fundamental to statistical learning. In specific, we like to highlight the following observations of statistical learning made in the paper; *Coordinates are rarely natural*, *Metrics are necessarily not justified* and *The need for large scale qualitative information*. The two first observations relate to the fact that as the dominant portion of statistical learning approaches work on vector spaces where the inner product is assumed to be naturally interpretable. However, observations are often "shoehorned" into vector spaces which are not natural in the sense that the inner product does not relate to the intrinsic structure of the data. In order to reason

about the space we require some form of similarity measure between points providing a distance or an ordering of the space. If the data is represented in vectorial space the natural similarity measure is the use of a norm. However, if the vectorial representation per say is not a natural representation of the data neither will the distance be. Especially relationships at large scale are likely to be less informative compared to local. This is indicated by the success of approaches which relaxes the assumption about the parametrisation to only assume it to be locally metric such as simple nearest neighbor methods [33, 34, 35] and the success of kernel induced feature spaces based on radial basis functions which emphasizes the local structure in the data. This is also the foundation for the last intuition that we wish to highlight from [32] that of the need for a qualitative measure of the data.

We have throughout this paper argued the importance of understanding the global structure of data. Given that it is only at best on a local scale we can associate significance to the similarity measure, we need tools that can in a principled manner provide qualitative measure on the global structure of a set of data induced by a local measure. A set of data and its structure can be studied by creating a graph where a node represents each samples with paths connecting nodes according to some similarity measure. Assuming that, we can at least on a local scale derive a somewhat natural notion of similarity, this graph represents the structure of the whole dataset that is induced by this local measure. The field of algebraic topology defines a formalism for providing qualitative measures on such graphs. However, one central question remains: on what scale the local similarity measure is relevant? In order to reduce the effects of noise in the samples we wish to use as large range of interaction as possible, however if too large we run the risk of connecting non-related components. This problem is well known in machine learning for constructing local affinity matrices [21, 36, 37]. In order to circumvent this problem the idea of Persistent Homology has been introduced which studies how the qualitative measure changes by varying the range of the local interactions. Persistent homology provides tools which can potentially make algebraic topology applicable as a formalism for studying uncertain data.

We believe that a symbiosis between statistical learnings tools with its principles for modeling in scenarios with uncertainty and missing data together with the tools for qualitative measurements of structure provided by topology has the potential of achieving a synergic effect for merging local observations and global structure in a unified framework.

7 What next?

Robots acting and interacting in realistic environments rely on perception, planning and control for motion generation. Although state of the art algorithms are capable of finding solutions that results in sucessfull goal generation in some applications, they are still not able to flexibly make use of the gathered experience and use it for solving a similar/related problem on a future occasion. Extracting the semantics of the task is one of the major bottlenecks that still remain to be solved and we argued

in thsi paper that this is in general dependent on using the *right* representation for the problem at hand. A good representation of data is one that except for being robust is capable of generalizing at the desired level.

In regard to motion generation, the classical approach operates in a complete configuration or state space represented at the level of generalized coordinates considering all joint angles and their 3D pose. This requires a computationally expensive state space optimization and randomized exploration in very large search spaces. In a EU funded project TOMSY (www.tomsy.eu) we study representations of actions and morphologies using topology-based abstractions in a layered manner and to implement dexterous manipulation on articulated and flexible objects using mappings between the topology-based abstract space, task space and joint space of metamorphic manipulators.

In this paper, have argued that one important mode of information for many application scenarios is not the actual variation in the data but the rather the higher order statistics as the structure of variations. We have exemplified this through a set of applications and show different ways of representing the structure of data, considering applications such as scene understanding, object recognition and data representation for grasping.

References

1. R. Rensink, J. ORegan, J. Clark, On the failure to detect changes in scenes across brief interruptions, Visual Cognition 7 (1) (2000) 127–145.
2. D. J. Simons, C. F. Chabris, Gorillas in our midst: Sustained inattentional blindness for dynamic events, Perception 28 (1999) 1059–1074.
3. D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.
4. B. D. Argalla, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robotics and Autonomous Systems 57 (5) (2009) 469–48.
5. V. Kruger, D. Kragic, A. Ude, C. Geib, The meaning of action: A review on action recognition and mapping, Advanced Robotics 21 (13) (2007) 1473–1501.
6. E. Aksoy, A. Abramov, F. Wörgötter, B. Dellen, Categorizing Object-Action Relations from Semantic Scene Graphs, in: IEEE International conference on robotics and automation, 2010, pp. 398–405.
7. I. Laptev, P. Perez, Retrieving actions in movies, in: IEEE International Conference on Computer Vision., 2007, pp. 1–8.
8. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
9. H. Kjellström, J. Romero, D. Kragic, Visual object-action recognition: Inferring object affordances from human demonstration, Computer Vision and Image Understanding 115 (2011) 81–90.
10. G. Luo, N. Bergström, C. H. Ek, D. Kragic, Representing Actions with Kernels, in: International Conference of Inteligent Robots and Systems, 2011.
11. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, Text classification using string kernels, The Journal of Machine Learning Research 2 (2002) 419–444.
12. N. Cristianini, J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge university press, 2006.
13. V. Kruger, D. L. Herzog, Sanmohan, A. Ude, D. Kragic, Learning actions from observations, Robotics and Automation Magazine 17 (2) (2010) 30–43.

14. D. Comaniciu, P. Meer, Mean Shift: A Robust Approach Toward Feature Space Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.
15. Y. Boykov, M.-P. Jolly, Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images, in: IEEE International Conference on Computer Vision, 2005, pp. 105–112.
16. N. Bergström, C. H. Ek, M. Björkman, D. Kragic, Scene Understanding through Interactive Perception, in: International Conference on Vision Systems, 2011.
17. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2010 (VOC2010) (2010).
18. R. Rusu, N. Blodow, M. Beetz, Fast Point Feature Histograms (FPFH) for 3D Registration, in: International conference on robotics and automation, 2009, pp. 3212–3217.
19. R. Rusu, A. Holzbach, N. Blodow, M. Beetz, Fast geometric point labeling using conditional random fields, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 7–12.
20. R. B. Rusu, Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments, Ph.D. thesis, Technische Universität München (2009).
21. J. B. Tenenbaum, V. de Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 290 (5500) (2000) 2319–2323.
22. Y. Teh, M. Jordan, M. Beal, D. Blei, Hierarchical dirichlet processes, Journal of the American Statistical Association 101 (476) (2006) 1566–1581.
23. J. Pitman, Combinatorial Stochastic Processes, St. Flour Summer School, Berlin: Springer-Verlag, 2006.
24. H. Wallach, S. Jensen, L. Dicker, K. Heller, An Alternative Prior Process for Nonparametric Bayesian Clustering, International Conference on Artificial Intelligence and Statistics.
25. R. Adams, H. Wallach, Z. Ghahramani, Learning the Structure of Deep Sparse Graphical Models, in: International Conference on Artificial Intelligence and Statistics, 2010.
26. T. L. Griffiths, Z. Ghahrmani, Infinite latent feature models and the Indian buffet process, in: Advances in Neural Information Processing, 2006, pp. 475–482.
27. D. Song, K. Huebner, V. Kyrki, D. Kragic, Learning Task Constraints for Robot Grasping using Graphical Models , IEEE/RSJ International Conference on Intelligent Robots and Systems (2010) 1579–1585.
28. C. H. Ek, D. Song, D. Kragic, Learning Conditional Structures in Graphical Models from a Large Set of Observation Streams through efficient Discretisation, in: International Conference on Robotics and Automation, Workshop on Manipulation under Uncertainty, 2011.
29. D. Song, C. H. Ek, K. Huebner, D. Kragic, Embodiment-Specific Representation of Robot Grasping using Graphical Models and Latent-Space Discretization, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011, pp. 1–8.
30. D. Song, C. H. Ek, K. Huebner, D. Kragic, Multivariate Discretization for Bayesian Network Structure Learning in Robot Grasping, in: International Conference on Robotics and Automation, 2011.
31. M. Titsias, N. Lawrence, Bayesian Gaussian Process Latent Variable Model, in: International Conference on Artificial Intelligence and Statistics, 2010.
32. G. Carlsson, Topology and data, American Mathematical Society 46 (2) (2009) 255–308.
33. G. Shakhnarovich, T. Darrell, P. Indyk, Nearest-neighbor methods in learning and vision, MIT Press, 2005.
34. G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, in: IEEE International Conference on Computer Vision, 2003, pp. 750–757.
35. O. Boiman, E. Shechtman, M. Irani, In defense of Nearest-Neighbor based image classification, in: Computer Vision and Pattern Recognition, 2008, pp. 1–8.
36. K. Q. Weinberger, F. Sha, L. K. Saul, Learning a kernel matrix for nonlinear dimensionality reduction, in: International Conference on Machine Learning, 2004.
37. S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science (290).