Project Acronym:	GRASP
Project Type:	IP
Project Title:	Emergence of Cognitive Grasping through Introspection, Emulation and Surprise
Contract Number:	215821
Starting Date:	01-03-2008
Ending Date:	28-02-2012

Deliverable Number:	D26
Deliverable Title :	Design and evaluation of representations and ontologies
Type (Internal, Restricted, Public):	PU
Authors	D. Kragic, D. Song, J. Bohg, T. Feix, J. Romero, T. Asfour, R. Dillmann, M. Do
Contributing Partners	KTH, OB, KIT

Contractual Date of Delivery to the EC:	28-02-2012
Actual Date of Delivery to the EC:	28-02-2012

Contents

Chapter 1

Executive summary

Deliverable D26: *Representations and Ontology for learning and Abstraction of Grasping*, presents the fourth year developments within WP2. According to the Technical Annex, deliverable D26 presents the activities in the context of Tasks 2.1-2.3:

- [**Task 2.1**] - Definition of the ontology: definition of sensory-motor control for action and object-action learning
- [**Task 2.2**] - Vocabulary of human and robot actions/interactions
- [**Task 2.3**] - Evaluation of representation: Evolving ontology through modeling of the perception-action cycle

The work in this deliverable relates to the following third year Milestones:

- [**Milestone 10**] - Linking structure, affordances, actions and tasks; evaluation of representations defined by the ontology.
- [**Milestone 11**] - Integration and evaluation of scenarios on multiple experimental platforms, demonstration of cognitive capabilities of robots.

The progress in WP2 is presented in the below summarized scientific publications, attached to this deliverable.

- In Attachment A, we continued to study embodiment-specific robot grasping tasks, represented in a probabilistic framework. The framework consists of a Bayesian network (BN) integrated with a novel multi-variate discretization model. The BN models the probabilistic relationships among tasks, objects, grasping actions and constraints. The discretization model provides compact data representation that allows efficient learning of the conditional structures in the BN. To evaluate the framework, we use a database generated in a simulated environment including examples of a human and a robot hand interacting with objects. The results show that the different kinematic structures of the hands affect both the BN structure and the conditional distributions over the modeled variables. Both models achieve accurate task classification, and successfully encode the semantic task requirements in the continuous observation spaces. In an imitation experiment, we demonstrate that the representation framework can transfer task knowledge between different embodiments, therefore is a suitable model for grasp planning and imitation in a goal-directed manner.
- In Attachment B, we study the representation problems in the context of high-dimensional data in particular. Many tasks in robotics and computer vision are concerned with inferring continuous or discrete state variables from observations and measurements of the environment. Due to the high-dimensional nature of the input data, inference is often approached in a two stage process: first a low-dimensional feature representation is extracted onto which secondly a learning algorithm is applied. Due to the significant progress that have been made 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. In Attachment B, we argue that in 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. We further argue that an important mode of information in many application scenarios is not the actual variation in the data but rather higher order statistics as the structure of variations. We exemplify this through a set of applications and show different ways of representing the structure of data.

- In Attachment C, motivated by the recent work on contextual recognition and estimation, we present a method for estimating the pose of human hands, employing information about the shape of the object in the hand. Despite the fact that most applications of human hand tracking involve grasping and manipulation of objects, the majority of methods in the literature assume a free hand, isolated from the surrounding environment. Occlusion of the hand from grasped objects does in fact often pose a severe challenge to estimation of hand pose. In the presented method, object occlusion is not only compensated for, it contributes to the pose estimation in a contextual fashion; this without an explicit model of object shape. Our hand tracking method is non-parametric, performing a nearest neighbor search in a large database (100 000 entries) of hand poses with and without grasped objects. The system operates in real time, is robust to self occlusions, object occlusions and segmentation errors, and provides full hand pose reconstruction from monocular video. Temporal consistency in hand pose is taken into account, without explicitly tracking the hand in the high-dim pose space. Experiments show the non-parametric method to outperform other state of the art regression methods, while operating at a significantly lower computational cost than comparable model-based hand tracking methods.

- Attachment D, we develop methods for evaluation of robotic and prosthetic hands capabilities where the human hand serves as a benchmark. In the design of hand prostheses, an open question is which degrees of freedom to actuate in order to achieve the best functionality of the hand. In robotics, apart from the actuation, the goal is also to develop highly dexterous hands. A natural question is how to define a similarity measure through which the capabilities of different hands can be analyzed. Many parameters can be taken into account - ranging from kinematic and dynamic properties to the choice of material (rigid vs. soft) and interaction with objects. Currently, there are no analytic methods for performing such analysis and the mainstream approaches perform exhaustive experimental evaluation. In this paper, we address the problem of comparing the capabilities of different hands through the use of non-linear dimensionality reduction techniques. We concentrate on the kinematic analysis - that is, we address the problem of how many different grasp types or how large space of poses different kinematic structures can achieve. In our study, we first generate data with human subjects, thus using the capabilities of the human hand as the benchmark. The generated human data is based on an extensive grasp taxonomy, including most common grasp types. We develop a methodology for comparing different anthropomorphic robotic and prosthetic hands for the specific task of object grasping. We show how different robotic hands perform with respect to the human hand, resulting also in a comparison between different robotic hand designs. Although the method is applied to hand data, it can be used to compare other types of kinematic structures as well.

- Attachment E, deals with the problem of observation and analysis of human motion that is often used for planning and control of human inspired movements in robots. Human data is usually high-dimensional and in many cases it is used to control a robot which much fewer degrees of freedom. To that end, different representations based on dimensionality reduction techniques have been used to enable viable control solutions. In control and planning of grasping movements in particular, postural synergies have been used as a low-dimensional representation to enable establishing correspondence between human and robot hand activities. In their original formulation, postural synergies are based on linear dimensionality reduction methods that, as we will show in this paper, do not represent human hand activity with sufficient accuracy due to inherit non-linearities in the data. Thus, the work presented in this paper addresses non-linear dimensionality reduction methods and their application to human hand data. In addition to adressing encoding of postural synergies, our work relates closely to recent work in robotic control of combined reaching and grasping movements. However, this work is based on an assumption that correlations in the data is evidence of causal relation, an assumption that may not hold. Non-linear dimensionality reduction methods may be used to tackle the correlations problem not by considering causal relations between dimensions, but by considering them being generated from an external manifold which has to be inferred. Showing how this can be done is the first contribution of our work. Another strong contribution of this paper is the analysis of the internal parameters used in dimensionality reduction techniques, which sheds light into algorithms which have been traditionally used as a black-box in robotics. Finally, we provide a thorough experimental evaluation that shows how the proposed methods outperform the standard techniques in the field both in terms of recognition and generation of motion patterns.

Appendix A

Attached Papers

[A] **Embodiment-Specific Representation of Robot Grasping using Graphical Models and Latent-Space Discretization**; Dan Song, Carl Henrik Ek, Kai Huebner and Danica Kragic; IEEE/RSJ International Conference on Intelligent Robots and Systtems, IROS 2011, San Francisco

[B] **The importance of structure**; Carl Henrik Ek and Danica Kragic; International Symposium of Robotic Research, ISRR 2011, Flagstaff

[C] **Non-Parametric Hand Pose Estimation with Object Context**, Javier Romero, Hedvig Kjellstrom, Carl Henrik Ek, Danica Kragic; Image and Vision Computing, (submitted)

[D] **Visualization of Anthropomorphic Hand Performance**; Thomas Feix, Javier Romero, Carl Henrik Ek, Heinz-Bodo Schmiedmayer, Danica Kragic; IEEE Transactions on Robotics (submitted)

[E] **Extracting Postural Synergies for Grasping** ; Javier Romero, Thomas Feix, Carl Henrik Ek, Hedvig Kjellstrom and Danica Kragic; IEEE Transactions on Robotics (submitted)

Embodiment-Specific Representation of Robot Grasping using Graphical Models and Latent-Space Discretization

Dan Song, Carl Henrik Ek, Kai Huebner and Danica Kragic

Abstract— **We study embodiment-specific robot grasping tasks, represented in a probabilistic framework. The framework consists of a Bayesian network (BN) integrated with a novel multi-variate discretization model. The BN models the probabilistic relationships among tasks, objects, grasping actions and constraints. The discretization model provides compact data representation that allows efficient learning of the conditional structures in the BN. To evaluate the framework, we use a database generated in a simulated environment including examples of a human and a robot hand interacting with objects. The results show that the different kinematic structures of the hands affect both the BN structure and the conditional distributions over the modeled variables. Both models achieve accurate task classification, and successfully encode the semantic task requirements in the continuous observation spaces. In an imitation experiment, we demonstrate that the representation framework can transfer task knowledge between different embodiments, therefore is a suitable model for grasp planning and imitation in a goal-directed manner.**

I. INTRODUCTION AND CONTRIBUTIONS

An important challenge in imitation learning [1] is the *correspondence problem* [2] due to the differences in embodiments between the teacher and the learner. Namely, direct copy of the demonstrated action may fail to achieve the goal of the demonstrated task, or even may not be feasible because the robot has different mechanical constraints. Several works have addressed the correspondence problem by constraining the imitation at a task space that is shared by the teacher and the learner. This common space can be either pre-specified by the user [3], or automatically identified using machine learning techniques [4]. In relation to robot arm movements, such a common space is usually the trajectory of the hand position and orientation in the Cartesian space, which is then reproduced by the robot solving the inverse kinematics [3].

However, identifying a common task space is difficult in the domain where the robot has to interact with the world: to grasp and manipulate objects. We may ask: *What is the common task space for* ***pouring*** *water into a cup?* Here, the robot has to consider not only the hand pose, finger configuration, but also the pose of the object, and its physical properties that determine if the object *affords* this task. Also, to firmly grasp an object for further manipulation, important control parameters such as the grasping force have to be considered. For the specific example of pouring, a good grasp would be the one that results in a stable manipulation of the objects during pouring, taking into account that the grasp should not be at a position so that the opening part is blocked.

To parameterize such semantic task constraints in a deterministic manner is hard. First, the task requirements can vary a lot with the task itself. For example, the constraint of a *hand-over* task is to leave enough free-space on the object so that it allows re-grasp. It is clearly described by a set of object and action variables that are different from those that define the pouring task. In addition, this task description may also be hand-specific. For example, human can apply power grasps to hand-over an apple, but a robot may fail with the same grasp type simply because it has a larger hand.

Our previous work [5] addressed already some of these challenges. We used a probabilistic graphical model – Bayesian Network (BN) – to encode such semantic task requirements for robot grasping. The network modeled the conditional distributions among a set of object and grasp related features that are hand-specific, together with the task requirements that have been introduced by human labeling. The initial results were very promising: the model allowed the robot not only to reason about high-level task representations, but also to make detailed decisions about which object to use and which grasp to apply in order to fulfill the requirements of the assigned task. However, the BN used in [5] models both discrete and continuous variables, which presents some limitations particularly in structure learning of the network. We therefore developed a novel multivariate discretization model presented in [6]. The model uses a non-linear dimensionality reduction technique to learn a low-dimensional latent representation of the observations. A mixture model is then learned to discretize the data allowing for a compact, generative representation of the data. The model is fully probabilistic and capable to facilitate structure learning from discretized data, while retaining the continuous representation.

The contribution of this paper is to create a fully probabilistic framework for embodiment-specific representation of robot grasping tasks. We do this by integrating the BN approach from [5] with the multi-variate discreitization model from [6]. The proposed approach is evaluated using human and robot object grasping examples in a simulated framework. We show that the two hands result in rather different network structures indicating potentially different conditional dependencies among the same set of task variables. Also, the conditional distributions in the individual variables turn to be hand specific. However, both models achieve good task classification, and represent the semantic task requirements in the continuous observation spaces.

D. Song, C.H Ek, K. Huebner and D. Kragic are with KTH – Royal Institute of Technology, Stockholm, Sweden, as members of the Computer Vision & Active Perception Lab., Centre for Autonomous Systems, www: http://www.csc.kth.se/cvap, e-mail addresses: {dsong,chek,khubner,danik}@csc.kth.se.

In an imitation experiment, we demonstrate that the proposed framework successfully transfers task knowledge between different hands and provides the means for grasp planning and imitation in a goal-directed manner. Compared with [5], [6], the current work extends the learning domain to a slightly more challenging dataset with more tasks, objects and embodiments.

II. Models

A Bayesian Network is a directed graphical model which exploits conditional dependencies in the data in order to learn an efficient factorization of the joint distribution in the data,

$$p(X_1, \ldots, X_N) = \prod_{i=1}^{N} p(X_i|\pi_i), \tag{1}$$

where X_i represents variables and π_i its parents in the network. The model is defined by a set of *parameters* defining each conditional model and by the *structure* of the vertices representing the conditional dependencies. Learning both structure and parameters from both continuous and discrete data poses a significant challenge. Most algorithms for structure learning only work with discrete variables therefore a pre-discretization step is necessary [7].

In [6] we developed a method capable of learning an intermediate discrete representation of a high-dimensional, continous observation space. In specific we apply techniques from generative dimensionality reduction – the Gaussian Process Latent Variable Model (GP-LVM) [8]. Its sparse variational formulation [9] provides both efficient learning of the latent space and the initial clusters for the subsequent discretization. Due to space limit, we refer the readers to [8], [9], [6] for detailed formulations of sparse GP-LVMs.

In this paper we improve the discretization model by incorporating an additional prior that encourages the location of the states to be sparse. In other words, we want a representation where each of the cluster centers are well separated in the latent space. To do so, we propose a prior over the discretization centers $\mathbf{U} = \{u_1, u_2, \ldots, u_M\}$, which are the inducing points of the sparse GP-LVM. This prior penalizes the L_1 norm of the off-diagonal elements in the inner-product matrix computed between the inducing points,

$$p(\mathbf{U}|\theta_U, \beta_U) = \mathcal{N}(\sqrt{D(\mathbf{U}, \theta_U)}|0, \beta_U^{-1}), \tag{2}$$

$$D(\mathbf{U}, \theta_U) = \sum_{ij}^{M} \lambda_{ij} k_u(u_i, u_j, \theta_U), \lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}.$$

If $k_u(u_i, u_j)$ is a smooth monotonically decreasing function with respect to $||u_i - u_j||$ the distribution will encourage a representation with well separated clusters. The parameters β_U and θ_U control the strength of the prior and the smoothness of k_u respectively. Here we use a radial basis function where θ_U controls the width of the function that also relates to the strength of the prior. Including the above prior into the method presented in [6] we are able to further improve previous results.

Once we have acquired a discrete version of the observations, we use a greedy search algorithm to find the structure, or the directed acyclic graph (DAG), in a neigborhood of graphs that maximizes the network score (Bayesian information criterion [10]). The search is local and in the space of DAGs, so the effectiveness of the algorithm relies on the initial DAG. As suggested by [11], we use another simpler algorithm, the maximum weight spanning tree [12], to find an oriented tree structure as the initial DAG. We assume the task class variable is the 'cause' of the systems thus the root node of the network.

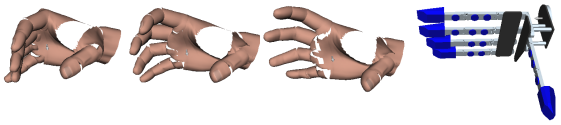


Fig. 1. Randomly sampled Eigengrasp preshapes of the human hand, and the preshape of Armar hand.

Inference

A trained BN defines an efficient factorization of the joint distribution of the observations. By converting the acyclic graph into a tree, the junction tree algorithm [13] allows efficient inference on the marginal distribution of any variable(s) conditioned on observations of others. The output of the inference is a multinomial distribution for variable X_i over each of its discrete states u_{ik} while the observation of the rest of the network $\mathbf{V}_i$ is at the state $\mathbf{v}_j$,

$$\mu_{ijk} = p(X_i = u_{ik}|\mathbf{V}_i = \mathbf{v}_j). \tag{3}$$

We will now describe how we can recover a continuous estimate of variable X_i in its original observation space $\mathbf{Y}$ from this distribution.

Each point of $\mathbf{x}_i$ on the latent space $\mathbf{X}$ defines a distribution over the observed data space $\mathbf{Y}$ through the GP that models the generative mapping. Therefore in order to acquire a continuous estimate in $\mathbf{Y}$ we need to determine a distribution over the latent space $\mathbf{X}$ associated with the multi-nominal distribution μ_{ijk}. In order to achieve this we first learn a parametric mixture model with the location of the inducing points as the mixture centers. In specific we use full-covariance Gaussian basis functions to define a mixture model with M components (discrete states),

$$p(\mathbf{x}_i) \propto \prod_{k=1}^{M} \lambda_k \mathcal{N}(\mathbf{x}_i|u_{ik}, \Sigma_{ik}), \tag{4}$$

and learn its parameters of the mixture model using the standard EM approach. The multinomial distribution output μ_{ijk} from the network defines a distribution over the inducing points u_{ik}. We use this distribution to specify the coefficient for the learned mixture components to create the following *conditional* mixture model,

$$p(\mathbf{x}_i|\mathbf{v}_j) \propto \prod_{k=1}^{M} \mu_{ijk} \mathcal{N}(\mathbf{x}_i|u_{ik}, \Sigma_{ik}). \tag{5}$$

We can then sample from the above distribution in order to find locations over the latent space which corresponds to our continuous estimate.

III. Data Generation

Tab. I shows the features used in this work. The features describing each grasp are divided into three sub-sets: *object features* (O) from the object representation, *action features* (A) from the planned grasps, and *constraint features* (C) resulting from the complementation of both, i.e. the hand-object configuration. Each grasp was visualized in GraspIt! to a human tutor who associated it with a task label (T).

Two hand models are used in the experiments: the human 20 degrees-of-freedom (DoF) hand, and the Armar 11 DoF hand [14]. The database includes in total 48 objects covering 6 object classes (8 models per class). Each object class includes 4 different object shapes each of which is scaled to 2 sizes – small and average. Four tasks are labeled: *hand-over, pouring, tool-use* and *dish-washing*. Compared to previous work we include the new task, *dish-washing*. In summary, the current approach extends [5], [6] with a more challenging dataset and a new robot hand showing the scalability of the framework.

Note that the human hand has an Eigengrasp pre-configuration $egpc$ as one of the action variables, whereas the Armar hand does not. Human hand is high-dimensional, but not all of the DoFs are indepently controlled. Therefore we use the idea of [15] to define random preshape configurations of the hand in the 2D *eigen grasp* space (i.e. $egpc$). The two dimensions of $egpc$ roughly represent the levels of finger spreading and finger flexion respectively. A detailed implementation on $egpc$ can be found in [5]. For the Armar hand the spreading component is missing, and the four fingers opposing the thumb can only flex and extend. We therefore do not implement random samples in preshape configuration for Armar hand, and the hand always starts at a preshape while all the DoFs are at zero, i.e. the fingers are fully extended (see Fig. 1).

Fig. 2 shows the schematic of the data generation process. To extract those features for each hand, we first generate grasp hypotheses using the grasp-planner BADGr [16], and evaluate them as scenes of object-grasp configurations in a grasp simulator, GraspIt! [17]. BADGr includes extraction and labeling modules to provide the set of variables presented in Tab. I. The interested reader is referred to [5], [16] for more details on the feature extraction. We emphasize that the grasp representation does not have to be non-redundant, e.g. *cvex* and *ecce* are allowed variables to both represent object shapes. Such an "over-representation" of the featured variables allows us to use BNs to identify the importance of, and dependencies between these variables in the scenarios of robot grasping tasks.

TABLE I

Used features with their dimensionality D (for continuous) and number of states N after discretization.

	Name	D	N	Description
T	$task$	-	4	Task Identifier
O_1	$obcl$	-	6	Object Class
O_2	$size$	3	8	Object Dimensions
O_3	$cvex$	1	4	Convexity Value $[0, 1]$
O_4	$ecce$	1	4	Eccentricity $[0, 1]$
A_1	dir	4	15	Approach Direction (Quaternion)
A_2	pos	3	12	Grasp Position
A_3	$upos$	3	8	Unified Spherical Grasp Position
A_4	$fcon$	11/20	6	Final Hand Configuration (Armar/Human)
A_5	$egpc$	2	8	Eigengrasp Pre-Configuration (only Human)
C_1	coc	3	4	Center of Contacts
C_2	$fvol$	1	4	Free Volume

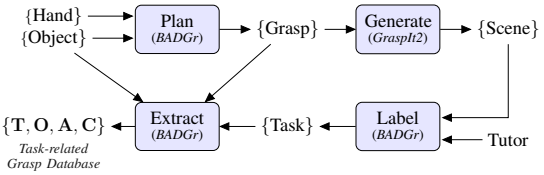


Fig. 2. Schematic diagram for generating task-related grasp database.

IV. Results

A. Experiment I: Structure Learning

The first experiment is to evaluate the network structure. Fig. 3 shows the results of learned DAGs for Armar (left) and human (right) hands. We note that learning the structure from data reveals complicated relationships among these variables, which will otherwise be very difficult to encode by human experts. An initial inspection of the DAGs associated with the different hands confirm our intuitive notion of the dependency relations between the variables. For example, the three action features – dir, $upos$ and pos – are connected with each other because the unified spherical grasp position $upos$ is directly derived from the grasp position pos and the hand orientation dir with respect to the object. And the object class $obcl$ determines the three object features $ecce$, $size$ and $cvex$.

We also noticed significant differences in the conditional structures between the two hand models. For instance, Armar hand has pos directly conditioned on $ecce$, whereas human hand does not. The reason might be that the Armar hand has limited kinematics configuration, therefore, when the object is quite eccentric, most stable grasps will have to be generated in the position around the side of an eccentric object, for example, on the handle of a hammer.

Also for human hand, center of contact coc has two parents $task$ and $obcl$, these links are both missing in the Armar hand. This is again explainable when consider the embodiment difference of the hands. The human hand has much more DoFs, and more flexible control in its pre-configuration (the random samples in the 2D Eigengrasp space $egpc$). This allows much more variation in its finger contacts with the object compared with those from Armar hand. As a result, coc which quantified this richer variation allows the learning algorithm to discover its potential relations with the object categories and the task requirements. Similar arguments also apply to the differences in connections around $fvol$, and $fcon$.

B. Experiment II: Task Classification

In this section we evaluate the learned network by their task classification performance. The performace is evaluated based on the testing data that also covers all the object

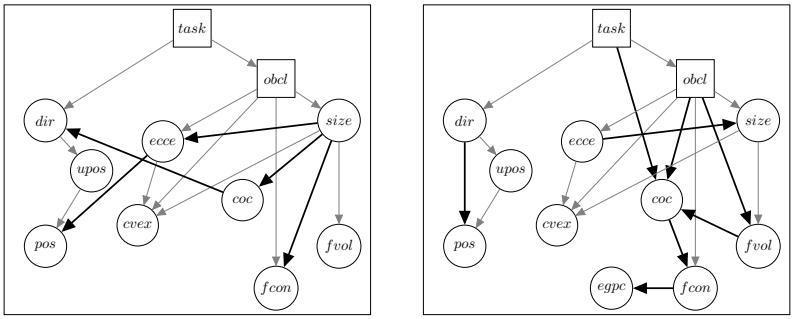


Fig. 3. Experiment I: The resulting DAGs by applying structural learning on (left) Armar hand, (right) human hand data. The differences in DAGs are highlighted by thick arrows. Square nodes represent discrete variables and circled nodes continuous.

	O	A	C	O, A	O, C	A, C	O, A, C
Armar Hand	0.00 0.32 0.33 0.35 0.00 0.46 0.00 0.54 0.00 0.00 1.00 0.00 0.00 0.00 0.13 0.87	0.15 0.09 0.33 0.43 0.02 0.63 0.02 0.33 0.00 0.30 0.59 0.11 0.02 0.07 0.13 0.78	0.00 0.02 0.31 0.67 0.00 0.00 0.57 0.43 0.00 0.00 0.78 0.22 0.00 0.04 0.33 0.63	0.24 0.28 0.26 0.22 0.02 0.78 0.00 0.20 0.00 0.00 0.93 0.07 0.02 0.02 0.13 0.83	0.00 0.37 0.33 0.30 0.00 0.50 0.00 0.50 0.00 0.00 1.00 0.00 0.00 0.13 0.13 0.74	0.33 0.11 0.26 0.30 0.04 0.46 0.22 0.28 0.00 0.11 0.89 0.00 0.07 0.04 0.26 0.63	0.30 0.26 0.20 0.24 0.02 0.76 0.00 0.22 0.00 0.00 0.93 0.07 0.02 0.02 0.13 0.83
Human Hand	0.24 0.35 0.12 0.29 0.00 0.46 0.00 0.54 0.62 0.00 0.38 0.00 0.07 0.00 0.02 0.91	0.54 0.42 0.00 0.04 0.01 0.94 0.00 0.05 0.29 0.71 0.00 0.00 0.38 0.33 0.00 0.29	0.23 0.62 0.15 0.00 0.01 0.95 0.04 0.00 0.00 0.51 0.49 0.00 0.13 0.76 0.11 0.00	0.44 0.28 0.20 0.08 0.01 0.75 0.00 0.24 0.28 0.00 0.72 0.00 0.06 0.01 0.02 0.91	0.28 0.30 0.14 0.28 0.00 0.49 0.00 0.51 0.05 0.00 0.95 0.00 0.01 0.04 0.07 0.88	0.51 0.39 0.07 0.03 0.01 0.90 0.01 0.08 0.07 0.44 0.49 0.00 0.33 0.26 0.07 0.34	0.47 0.28 0.16 0.09 0.01 0.80 0.00 0.19 0.04 0.00 0.96 0.00 0.03 0.05 0.06 0.86

Fig. 4. Experiment II: Confusion matrices for task classification given different observations spaces: permutations of O, A, C features. For each 4×4 matrix, from left to right (top to down), the 3 tasks are: *hand-over, pouring, tool-use, dish-washing*.

classes. The data size is one quarter of the training cases.

As shown in Fig. 4, this task classification is based on the inference of $task$ variable given observation of different set of other variables that form a complete permutation of the 3 feature sub-sets: O, A and C. For object features O, we assume that the object is unknown, therefore object class information *obcl* is not observed. This is to simulate the real-world scenarios where recognizing object categories from its raw features is still a hard problem for robot sensor systems.

Comparing the task classification given different observation spaces (different columns), we see that for both hands, the object and action features (O, A) result in quite good task classification on the last 3 tasks: *pouring, tool-use* and *dish-washing*; particularly for the Armar hand, the accuracy are 78%, 93% and 83% respectively. When the two constraint features $fvol$ and coc are also observed (O, A, C), we observe overall improvements for human hand, but not so much for Armar. This can be explained by the differences in DAGs where Armar hand has less conditional dependencies discovered with the two constraint variables.

When only object features are observed, both hands have good classification on *dish-washing* task with slight confusion with *tool-use*. This is because in the labeled grasp data, the objects that are good for *dish-washing* are all the mugs and glasses, and one particular knife model (the kitchen knife). But the *pouring* task is never confused with *tool-use* because no tool objects affords pouring, and the observed object features could clearly differenciate the tools from the container objects. However, the grasps that are good for *pouring* is often confused with *dish-washing* becasue many pourable objects are also applicable to be dish-washed.

The *hand-over* task is often confused with others even when most features are observed (column O, A, C). This is expected as grasps that are good for *hand-over* are in many cases also likely to work well for the other three. This indicates that our classification of task might need a hierarchical structure rather than the flat class association we use here.

When comparing the confusion matrices between two hand models, we see that in any observation conditions, the performance over task classification has very different profiles in different hands. This means a variable that is strong in task description for one hand might be weak for another, again supporting the idea of embodiment-specific representation for grasping tasks.

C. Experiment III: Inference on Unified Grasp Position

In Experiment II we showed that the two hand BNs have different but good performances in task classification. The goal of Experiment III is to examine i) if both models could successfully encode task constraint in the continuous space of object observation, and ii) if this constraint is hand-dependent. Notice that the constraint of a given task is often encoded by a combination of multiple features, e.g. one should not grasp this object from this position pos, in this orientation dir, and with this joint configuration $fcon$. However due to space limit and for the purpose of

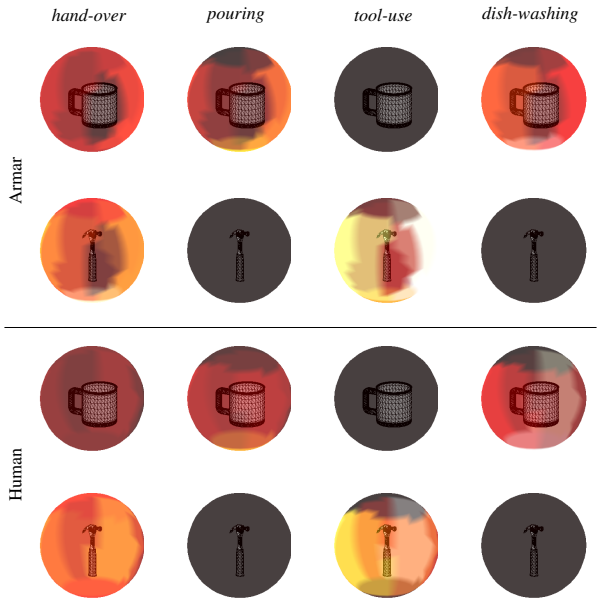


Fig. 5. Experiment III: Likelihood maps of the unified grasp position given tasks and object features $P(upos|T, O)$. The top panel is for Armar hand and bottom panel for human hand.

easier evaluation by the readers, we choose the unified grasp position $upos$ which combines the absolution grasp position pos and approach vector of the hand as an intuitive variable to visualize the task constraint.

For each hand, we sample 625 points on the unified sphere (where $upos$ is located) around the object. As shown in Fig. 5, for each sampled point, the likelihood is obtained given the 4 tasks, and the object features for 2 unknown objects: a mug and a hammer, i.e. $P(upos|T, O)$. The top panel shows the results for the Armar hand, and the bottom for the human hand.

We see that, for both hands, the model sucessfully rules out the mug for *tool-use*, and the hammer for *pouring* and *dish-washing* tasks. For *pouring*, the mug can not be grasped from the top as it will block the opening; similarly, when using the hammer as a tool, the grasp should avoid the head of the hammer as it is the functional part. To wash the mug, the preferred grasps indicated by the network are clearly from side or bottom. This is because the mugs usually need to be placed upside-down in the dishwasher, so grasping from top is not so convenient for this task.

When comparing the likelihood maps between the two hands, we have very interesting observations. In general maps are different for the two different embodiments even though they all model the task constraints in a similar way. In a specific case of *hand-over* the hammer, Armar hand has quite low likelihood on the side of the hammer that is facing the head of the hammer. Thinking closely, we understand that grasping from this position is particularly difficult for the Armar hand because the fingers might contact the sharp edges on the hammer head, resulting in unstable configuration. Similar situation is also for grasping from the top approaching the hammer head. On the contrary, human hand has much more uniformed distribution around the hammer.

This experiment again demonstrated the strength of the proposed framework: by modeling the embodyment-specific task space using a probabilistic network, we have learned not only the affordances of the objects based on its basic 3D features, but also the robot's own motor capability.

D. Experiment IV: Goal-directed Imitation

Finally we would like to demonstrate the application of the proposed framework in the scenarios of goal-directed imitation. The experiment is implemented using the human hand model as the demonstrator, and the Armar hand as the imitator. The goal is to imitate the demonstrator performing the *pouring* (demo 1) and *dish-washing* (demo 2) tasks using a mug (see Tab. II), and the *hand-over* (demo 3) and *tool-use* (demo 4) tasks using a hammer (see Tab. III). The object images in step 1 and step 2.1 shown in both tables are presented with same scale, so the size of the objects can be compared. We use $\mathbf{o}^H, \mathbf{a}^H, \mathbf{c}^H$ to indicate the human demonstrated object, action and constraint features respectively, $\mathbf{o}, \mathbf{a}, \mathbf{c}$ to represent the instances of the features of the Armar hand.

The process of the imitation consists of two major steps: step 1 for task recognition, and step 2 for object or action selection, the same way as we outlined in [5]. Briefly, in step 1, the robot uses the human hand-specific network to recognize the demonstrated task $\hat{t}^H$ based on maximum-log-likelihood estimation $L^H(\ t \mid \mathbf{o}^H, \mathbf{a}^H, \mathbf{c}^H)$, where L^H denote log-likelihood using human network. In step 2, given this recognized task $\hat{t}^H$ as the goal, the robot choose the object among the ones in the scene, and then select the most compatible grasp on the chosen object to achieve the task. Object and action selection has been formulated as the Bayesian decision problems, where a reward function is a weighted combination of their *task affordance* represented by the likelihood function L^R and the similarity to the demonstration S. The weight λ is a high-level control input to define the imitation requirements. Due to space limit, we refer the reader to [5] for the detailed formulation of the Bayesian decision problem and the confidence-based similarity metric.

Tab. II and III present the results of the imitation experiment. The bar plots on the right side of the tables show the log-likelihood values for step 1, and the reword functions in step 2.1 and 2.2. We see that in all four demonstrations, the robot could correctly recognize the tasks, even though there might be potential confusion with *hand-over* task (in demo 1 2 and 3). In demo 3, we find an interesting result where the grasping on the hammer has returned the zero probability for *tool-use*, and low but non-zero probability for *pouring* and *dish-washing*. Aparently unintuitive, but the result is consistent with what we have observed previously [5]: since we assumed unknown object, the inference was only based on observation of object $size$, $cvex$ and $ecce$ features, the network has confused a hammer with other container objects like bottles and glasses.

TABLE II

EXPERIMENT IV: GOAL-DIRECTED IMITATION ON *'pouring, dish-washing'* TASKS.

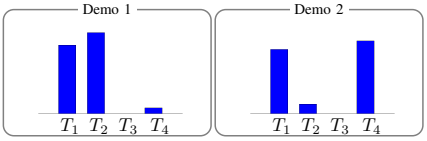
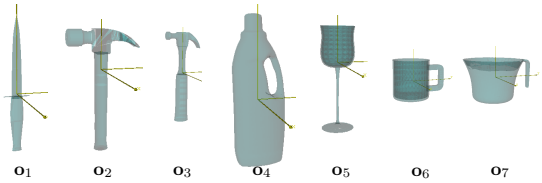
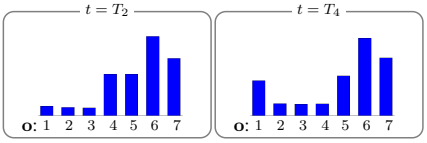
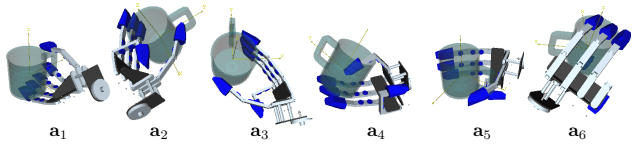
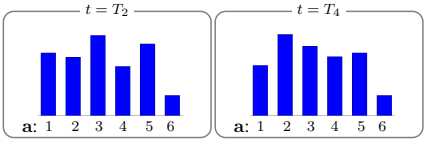
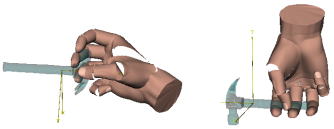
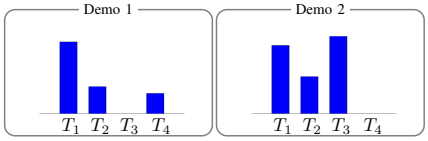
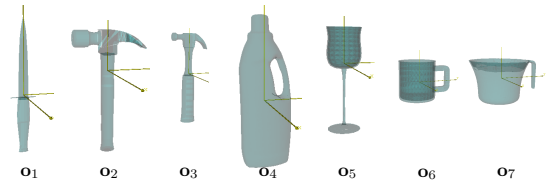
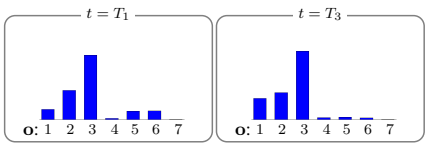
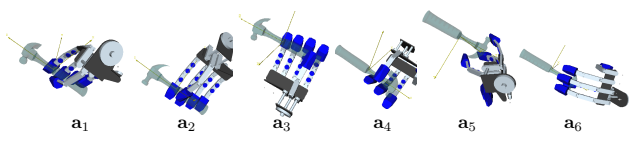
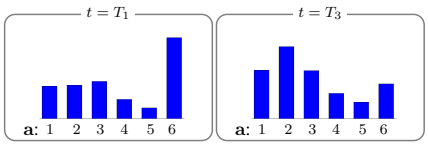
Scenes	Objective Functions
Step 1 Human demonstration: recognize task $\hat{t}^H$	$L^H(\,t \mid \mathbf{o}^H, \mathbf{a}^H, \mathbf{c}^H)$
 Demo 1 Demo 2 $T_1 = \textit{hand-over}$ $T_2 = \textit{pouring}$ $T_3 = \textit{tool-use}$ $T_4 = \textit{dish-washing}$	
Step 2.1 Select object $\mathbf{o}^*$: matching $\hat{t}^H$, and also similar to $\mathbf{o}^H$	$L^R(\,\mathbf{o} \mid t) \cdot 0.5 + \; S(\mathbf{o},\, \mathbf{o}^H \mid t) \cdot 0.5$
 $\mathbf{o}_1$ $\mathbf{o}_2$ $\mathbf{o}_3$ $\mathbf{o}_4$ $\mathbf{o}_5$ $\mathbf{o}_6$ $\mathbf{o}_7$	
Step 2.2 Select action $\mathbf{a}^*$: matching $\hat{t}^H$, and also similar to $\mathbf{a}^H$	$L^R(\,\mathbf{a} \mid t, \mathbf{o}_6) \cdot 0.5 + \; S(\mathbf{a},\, \mathbf{a}^H \mid t) \cdot 0.5$
 $\mathbf{a}_1$ $\mathbf{a}_2$ $\mathbf{a}_3$ $\mathbf{a}_4$ $\mathbf{a}_5$ $\mathbf{a}_6$	

TABLE III

EXPERIMENT IV: GOAL-DIRECTED IMITATION ON *'hand-over, tool-use'* TASKS.

Scenes	Objective Functions
Step 1 Human demonstration: recognize task $\hat{t}^H$	$L^H(\,t \mid \mathbf{o}^H, \mathbf{a}^H, \mathbf{c}^H)$
 Demo 3 Demo 4 $T_1 = \textit{hand-over}$ $T_2 = \textit{pouring}$ $T_3 = \textit{tool-use}$ $T_4 = \textit{dish-washing}$	
Step 2.1 Select object $\mathbf{o}^*$: matching $\hat{t}^H$, and also similar to $\mathbf{o}^H$	$L^R(\,\mathbf{o} \mid t) \cdot 0.5 + \; S(\mathbf{o},\, \mathbf{o}^H \mid t) \cdot 0.5$
 $\mathbf{o}_1$ $\mathbf{o}_2$ $\mathbf{o}_3$ $\mathbf{o}_4$ $\mathbf{o}_5$ $\mathbf{o}_6$ $\mathbf{o}_7$	
Step 2.2 Select action $\mathbf{a}^*$: matching $\hat{t}^H$, and also similar to $\mathbf{a}^H$	$L^R(\,\mathbf{a} \mid t, \mathbf{o}_3) \cdot 0.5 + \; S(\mathbf{a},\, \mathbf{a}^H \mid t) \cdot 0.5$
 $\mathbf{a}_1$ $\mathbf{a}_2$ $\mathbf{a}_3$ $\mathbf{a}_4$ $\mathbf{a}_5$ $\mathbf{a}_6$	

In step 2.1, the robot is able, in all four demonstrations, to choose among seven objects the one that matches the goal of the task $\hat{t}^H$ and at the same time is also similar to the object used by the human hand. In Tab. II we see the network preferred the smaller mug $\mathbf{o}_6$ that is similar size to the mug in the demonstration in both *pouring* and *dish-washing* tasks. In *dish-washing* task, the knife $\mathbf{o}_1$ has almost as high reward value as the glass $\mathbf{o}_5$. This is because one kitchen knife in the knife category affords *dish-washing*.

Finally in step 2.2, the robot successfully selected the grasp hypotheses that satisfy the requirements on task affordance and grasp similarity. In *pouring* task, grasp $\mathbf{a}_6$ has lowest ranking, which is obvious as three fingers block the cup opening. Grasp $\mathbf{a}_5$ is a very natural configuration for the *pouring*. But it is ranked as the second best grasp because compared to $\mathbf{a}_3$, $\mathbf{a}_5$ is less similar to the demonstrated grasp. Similar behaviors have been observed in other 3 demonstrations.

V. Conclusion

We have proposed a unified probabilistic framework to represent the embodiment-specific grasping tasks. The framework consists of a discrete Bayesian network and the sparse GP-LVM-based multi-variate discretization method. The Bayesian network models the task constraint through conditional distributions among a set of task, object, action and constraint variables. The discretization model provides compact, efficient data representations that allow fast learning and inference for the Bayesian network. With the simulated data from a human and a robot hand, we have shown that the grasping tasks are hand-specific, and the differences are reflected both in the conditional (in)dependencies between the representation variables (network structure), and in the probabilistic distributions of individual variables. However, both models perform well in task classification and representation of the underlying constraints.

We also showed that the hand-specific task representation can provide a unified framework for many aspects in scenarios of goal-directed grasp imitation. Not only can the robot recognize the intention of the human demonstration, but it can also reason in the low-level feature space of the object and grasp actions conditioned on the high-level task requirements. As a result, the robot can make automatic decisions that satisfy multiple user requests, for example, task affordance and grasp similarity.

Though in this paper, the proposed framework was only experimented with one grasp planner [16], we want to emphasize that it is not limited to any specific grasp planning systems. Several grasp planners can provide different representations of grasps and objects, and together with a human-provided task information, we could obtain similar task constraint models for each hand. In the cases the two grasp planners can provide similar grasp-related variables, we expect that the model trained on one planner could be used to infer task information on the other. This is to be tested in one of the next steps in the future research.

In addition, there are also some limitations in discretization model that need further research. Currently, the number of discrete states are manually chosen to satisfy a trade-off between refined data representation and complexity of BNs. In the future, we would also like to learn this hyper parameter automatically from data.

Finally, we plan to test this framework in grasp planning and execution in real robot platforms where sensorimotor uncertainty is more prominant. We believe this will further exemplify the benifits of using a probabilistic model capable of dealing with uncertainty in real-world applications.

Acknowledgments

This work was supported by EU IST-FP7-IP GRASP and Swedish Foundation for Strategic Research.

References

[1] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, "Robot Programming by Demonstration," in *Springer Handbook of Robotics*, 2008, pp. 1371–1394.

[2] C. L. Nehaniv and K. Dautenhahn, *The Correspondence Problem*. MIT Press, 2002, pp. 41–61.

[3] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009.

[4] S. Bitzer, I. Havoutis, and S. Vijayakumar, "Synthesising novel movements through latent space modulation of scalable control policies," in *the 10th International Conference on Simulation of Adaptive Behavior*, 2008.

[5] D. Song, K. Huebner, V. Kyrki, and D. Kragic, "Learning Task Constraints for Robot Grasping using Graphical Models," in *IROS*, 2010.

[6] D. Song, C.-H. Ek, K. Huebner, and D. Kragic, "Multivariate discretization for bayesian network structure learning in robot grasping," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2011, to appear.

[7] L. D. Fu, "A Comparison of State-of-the-art Algorithms for Learning Bayesian Network Structure from Continuous Data," Master's thesis, Biomedical Informatics, Vanderbilt University, December 2005.

[8] N. D. Lawrence, "Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, November 2005.

[9] M. Titsias, "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *Artificial Intelligence and Statistics*, 2009.

[10] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] P. Leray and O. Francois, "BNT Structure Learning Package: Documentation and Experiments," Université de Rouen, Tech. Rep., 2006.

[12] C. Chow and C. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[13] S. Lauritzen and D. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 157–224, 1988.

[14] T. Asfour, K. Regenstein, P. Azad, J. Schrder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "Armar-iii: An integrated humanoid platform for sensory-motor control," in *Proceedings of the 7th IEEE-RAS International Conference on Humanoid Robots*, 2006.

[15] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dexterous Grasping via Eigengrasps: A Low-dimensional Approach to a High-complexity Problem," in *RSS 2007 Manipulation Workshop*, 2007.

[16] K. Huebner, "BADGr - A Toolbox for Box-based Approximation, Decomposition and GRasping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems: Workshop on Grasp Planning and Task Learning by Imitation*, 2010, **URL:**http://www.csc.kth.se/∼khubner/badgr/.

[17] A. T. Miller and P. K. Allen, "GraspIt! A Versatile Simulator for Robotic Grasping," *IEEE Robotics and Automation Magazine*, 2004.

The importance of structure

Carl Henrik Ek and Danica Kragic

Abstract Many tasks in robotics and computer vision are concerned with inferring continuous or discrete state variables from observations and measurements of the environment. Due to the high-dimensional nature of the input data inference is often approached in a two stage process: first a low-dimensional feature representation is extracted onto which secondly a learning algorithm is applied. Due to the significant progress that have been made 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. In this paper we argue that in 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 an important mode of information in many application scenarios is not the actual variation in the data but rather 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 possible

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

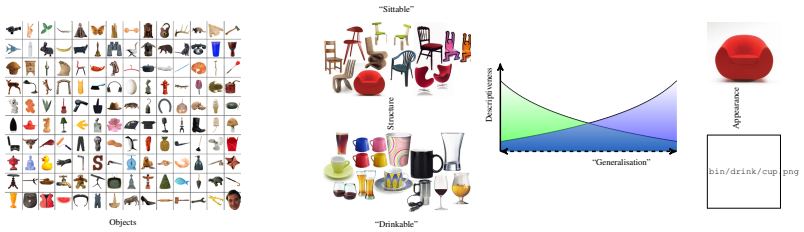


Fig. 1 *The above figure tries to highlight the notion of the importance of structure that we try to convey in this paper. The example above shows a large data-base of objects to the far left. Of these we want find a representation in order to classify objects at a certain resolution. If the representation 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 "sitable" 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.*

to design artificial systems whose 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 followed the same development. The 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 sampled from 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 strength of human inference is our capability of being selective with the information we use 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 categorisation 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 representations the discriminating information stands for only a small portion

of the variance. Such representations often implies a significant challenge in terms of modeling and inference.

In this paper we argue that rather than focusing on building models capable of representing larger portions of the variance in the sensory, we should aim to carefully consider what information that is actually relevant for the problem at hand. 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 for a specific problem but rather to exemplify a argument through a range of applications in order to stimulate further discussion on the topic.

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 task of object modeling. This provides an intuitive example of the concepts that we address in this paper. Object modeling is a 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 objects 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. 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.

Atatistical methoda rely 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. To that end the traditional approach have 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: *W*hat 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 fore-filled 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. This 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.

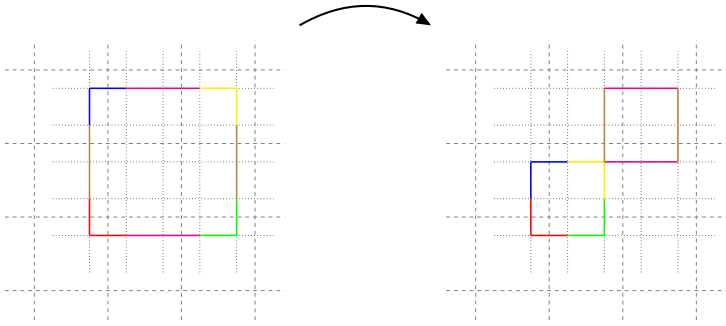


Fig. 2 *The above figure shows two different objects with two different scales of local representation, dotted (fine) and dashed (coarse). First order statistics from the fine resolution will not be able to discriminate the two objects while at the coarser scale they will be different. However, using 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 a larger amount of local features by acquiring large (and growing!) training data sets. The hope have been that by exploiting supervised machine learning techniques, such as kernel machines or metric learning, we can acquire a representation with the desired balance between generalization and discrimination.

We argue that 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 we believe are going to be useful providing 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

Robots often have to work 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 cutlery are cleared before the plates or vice-versa. In this section we describe different applications where we, through a model of temporal structure, manage to simplify an otherwise complicated inference task.

3.1 Interaction

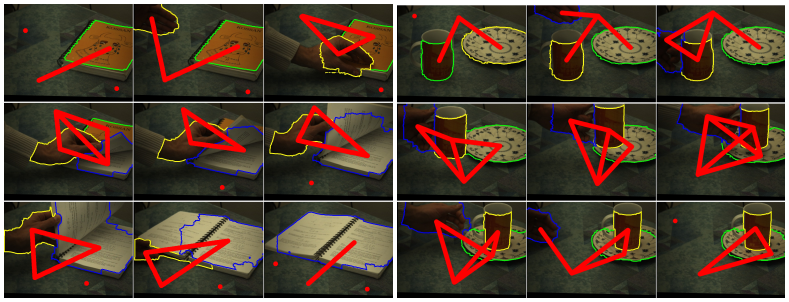


Fig. 3 *The* **left** *example shows an instance of the* **Opening Book** *action while the* **right** *shows the* **Moving Object**. *In each of the images the result of the segmentation and its corresponding graph have been overlaid . Only the spatial relations between the segments are extracted and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

Recently, [6] suggested a method for action classification by constructing an image feature 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 sequence. The temporal structure is encoded by a graph representing each frame, every object being a node and connected com-

ponent sharing an edge, see Fig 3. This process removes all information associated with appearance and identity leaving only the interaction between the objects. 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 from 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 to 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 parameterization 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. The inner product is defined as a function of the matching part of the overlap between two strings, see Fig 4. Clearly the space 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

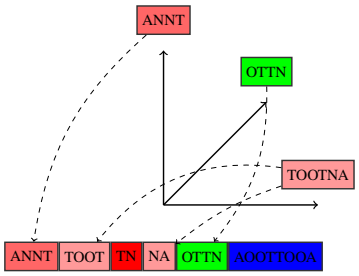


Fig. 4 *For a the specific semantic alphabet, here defining the four different interaction relationships between objects:* $\{A,N,T,O\}$*, we above show a subspace of the feature space representing the sequence. The sequence* **ANNT** *(red) and* **OTTN** *(green) exists in order in the string and will 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 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 rigid motion assumption. In [16] we describe an approach combining local appearance cues with a method for modeling rigid motion using 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.

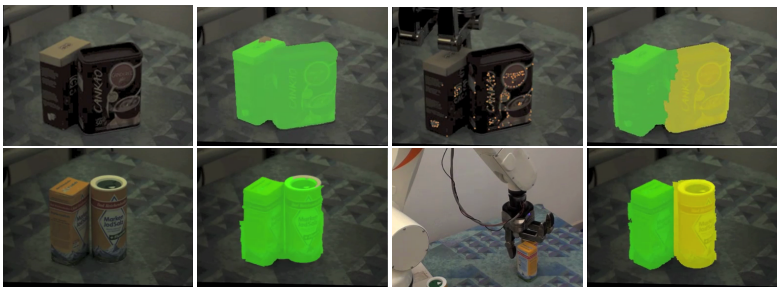


Fig. 6 *The left most column shows two scenarios where two objects have been placed on a table top. Using a traditional appearance based image segmentation approach it is not possible to separate the objects. By introducing motion in to the scene by letting the robot interact with the environment the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

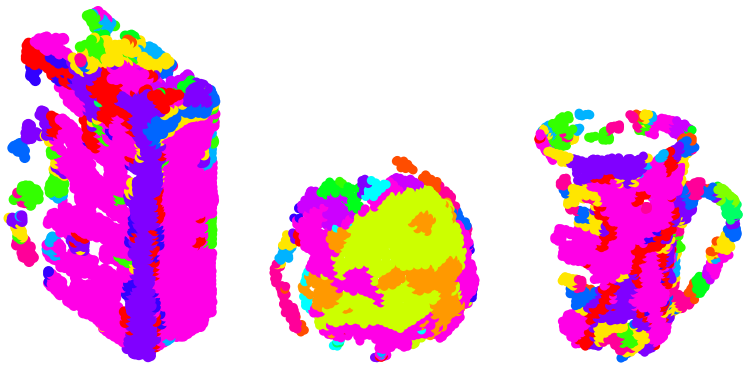


Fig. 7 *Object features representing the local geometrical class encoded by colour shown for three different objects, from the left box, citrus fruit and mug.*

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 patches.

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 categorisation 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 treating 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 yet. 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 discretization 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 discretization 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

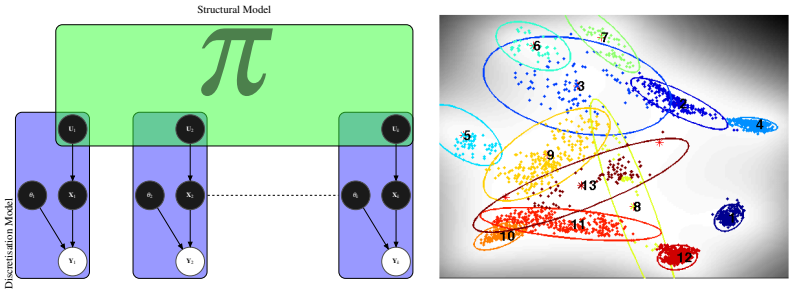


Fig. 8 *The left image shows a schematic graphical model of the structure learning approach. For each continuous observation space* $\mathbf{Y}_i$ *we learn a low dimensional representation* $\mathbf{X}_i$ *with a functional relationship to the observed data parametrised by* θ_i*. Further, the low-dimensional 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 discretization colour coded. The separation between the clusters is controlled by a prior modeling the trade-off between discretization 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 for discretization or in the object category example by on the local feature level ex-

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)

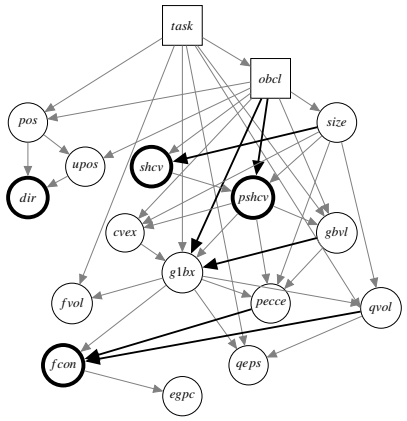


Fig. 9 *Example of a learned factorised representation of* 17 *different observation spaces for a grasping scenario. To the left the different features are shown and to the right the resulting graphical model with the learned structure. The structure is very complicated and it is highly unlikely that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

tracting 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 alleviate 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 pro-

viding 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 parameterization 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 learning 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 successful 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. We have argued in this 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).

Non-Parametric Hand Pose Estimation with Object Context

Javier Romero*, Hedvig Kjellström, Carl Henrik Ek, Danica Kragic

CVAP/CVAS, KTH, SE-100 44 Stockholm, Sweden

Abstract

In the spirit of recent work on contextual recognition and estimation, we present a method for estimating the pose of human hands, employing information about the shape of the object in the hand. Despite the fact that most applications of human hand tracking involve grasping and manipulation of objects, the majority of methods in the literature assume a free hand, isolated from the surrounding environment. Occlusion of the hand from grasped objects does in fact often pose a severe challenge to estimation of hand pose. In the presented method, object occlusion is not only compensated for, it *contributes* to the pose estimation in a contextual fashion; this without an explicit model of object shape. Our hand tracking method is non-parametric, performing a nearest neighbor search in a large database (100 000 entries) of hand poses with and without grasped objects. The system operates in real time, is robust to self occlusions, object occlusions and segmentation errors, and provides full hand pose reconstruction from monocular video. Temporal consistency in hand pose is taken into account, without explicitly tracking the hand in the high-dim pose space. Experiments show the non-parametric method to outperform other state of the art regression methods, while operating at a significantly lower computational cost than comparable model-based hand tracking methods.

Keywords: Articulated Hand Pose, Approximate Nearest Neighbor, Context

*Corresponding author.

Email addresses: jrgn@kth.se (Javier Romero), hedvig@kth.se (Hedvig Kjellström), chek@csc.kth.se (Carl Henrik Ek), dani@kth.se (Danica Kragic)

URL: www.csc.kth.se/~jrgn (Javier Romero), www.csc.kth.se/~hedvig (Hedvig Kjellström), www.csc.kth.se/~chek (Carl Henrik Ek), www.csc.kth.se/~danik (Danica Kragic)

An early version of this paper appears in [1].

1. Introduction

Human pose estimation is an important task for applications such as teleoperation and gaming, biometrics and prosthesis design, and human-robot interaction. However, accurate 3D reconstruction of human motion from images and video is a highly non-trivial problem, characterized by high-dimensional state spaces, fast and non-linear motion, and highly flexible model structures [2]. All this is applicable to hand reconstruction as well as full body reconstruction [3, 4, 5, 6]. However, while a full body pose estimator encounters additional challenges from e.g. clothing, a hand pose estimator has to deal with other but equally demanding issues: similarity in appearance between different parts of the hand (e.g. different fingers), and large self occlusion.

An important aspect of hand pose estimation is that humans are frequently holding objects. This is the case in the majority of the application areas mentioned above. The grasped object is often occluding a large part of the hand – for a plausible example, see Figure 1, left.

Despite this, researchers have up to now almost exclusively focused on estimating the pose of hands in isolation from the surrounding scene, e.g. [7, 8, 9, 10, 11]. As illustrated in Figure 1, top and middle, this will be inadequate if the observed hand interacts closely with objects during estimation.

Object-contextual hand pose estimation has been addressed in a generative manner in two recent works. In [12] the authors show that the hand pose can be reconstructed robustly despite the object occlusion. In [13], this is taken one step further, with explicit reconstruction of the object in 3D. By enforcing physical constraints on the hand pose from the object 3D surface and vice versa, the two pose estimation processes guide each other.

In contrast to [12, 13], we take a discriminative approach to object-contextual hand pose estimation. *The main contribution of this paper is a method for estimating human hand pose, employing contextual information about the shape of the object in the hand.* Neither the hand nor the object are explicitly reconstructed; the hand and the object are instead modeled together, encoding the correlations between hand pose and object shape in a non-parametric fashion. In spirit of recent methods for contextual recognition and estimation, e.g. [3, 14, 13, 6], the object occlusion thereby helps in the hand pose reconstruction.

There are two reasons for exploring discriminative hand pose estimation with object context. Firstly, while generative estimation approaches commonly are

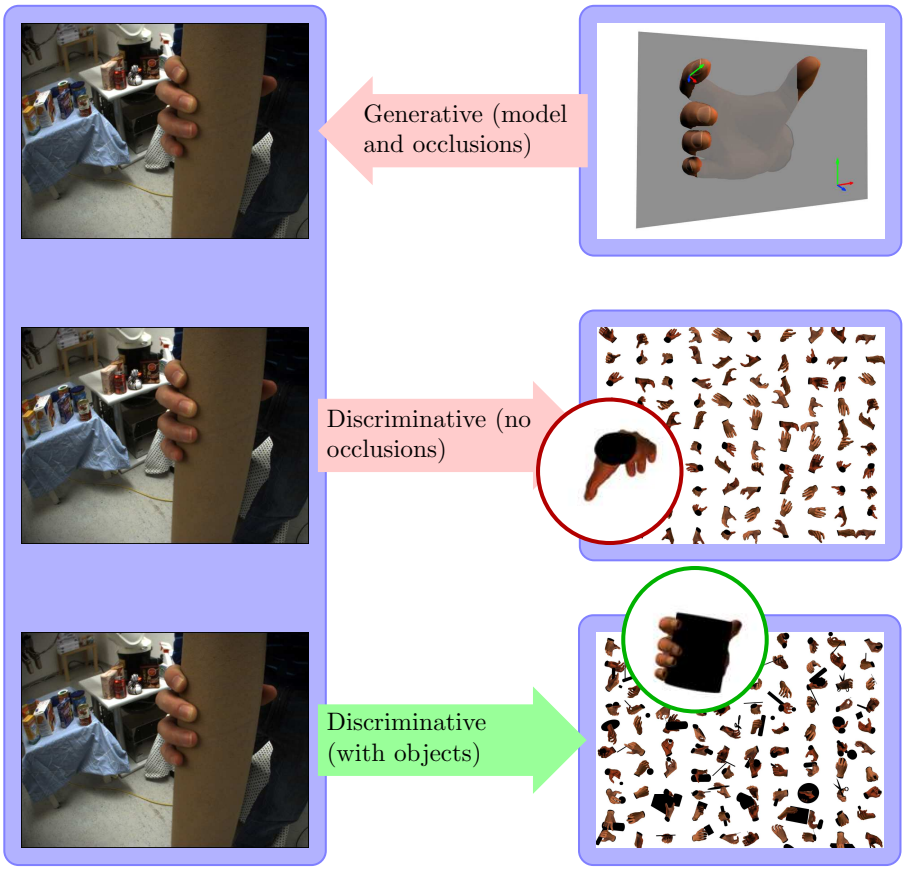


Figure 1: Hand pose estimation is traditionally approached in two different manners, either with a generative model (top) or using a discriminative approach (middle). With a generative model, a model of the hand is maintained, and the image of the model is evaluated against the observed image. In a discriminative approach, the image generation process is not explicitly modeled; instead, a (parametric or non-parametric) mapping from image to pose is learned from training examples. If objects are not taken into regard in the modeling process, both these approaches have significant problems predicting in scenarios where large portions of the hand are occluded. In the generative case (top), there is too little image evidence to compute an informative likelihood. In the discriminative case (middle), the learned mapping can not take the object occlusion into regard, and will return an erroneous estimate. Our method (bottom) addresses this problem, by exploiting contextual information in the scene such as object-hand interaction. Due to this we can reliably predict pose in scenarios with significant occlusion. We would like to point out that our model is not limited to scenarios where an object is being manipulated but equally valid to estimate a free hand. Objects can also be taken into regard in a generative framework; see Section 2.

more accurate, discriminative approaches are commonly more robust and computationally efficient; this is discussed further in Section 2. In, e.g., robotic applications, computational speed is critical, making discriminative approaches attractive. It is therefore valuable to investigate the possibility of estimating hand pose discriminatively in the context of objects.

Secondly, apart from the purely physical object constraints on the hand pose [13], there is also a functional correlation between object shapes and the manner in which they are grasped by a hand [15]. Thus, all physically possible ways of grasping an object are not equally likely to occur during natural object manipulation activities. Probability densities over hand pose conditioned on object shape can be encoded (in a non-parametric manner) in our discriminative method, while it is difficult to encode this information in a generative model based method.

Figure 1, bottom row illustrates our approach. In our non-parametric method, pose estimation essentially corresponds to matching an observed hand to a very large database (100 000 entries) of hand views. Each instance in the database describes the articulation and the orientation of the hand. The configuration of a new (real) image can then be found using an approximate nearest neighbor approach, taking previous configurations into account.

In our system, the database contains hands both with and without grasped objects. The database depicts grasping hands including occlusion from objects with a shape *typical for this kind of grasp*; this encodes functional correlations between object shape and the articulation of the grasping hand. The occlusion shape is strongly correlated to grasping type which further has a strong dependency with the hand articulation. Since the underlying assumption is that appearance similarity can be related to similarity in hand pose the object shape contributes to the hand pose estimation.

In many scenarios it is hard to differentiate between the palm and the dorsal ("back-hand") side of the hand. However, the object is much more likely to occlude the palm rather than the dorsal side of the hand. This is an example of how object knowledge can be exploited in order to resolve the ambiguities typically associated with hand pose estimation.

The rest of the paper is organized as follows: In Section 2 the relations to related work are discussed. The probabilistic estimation framework is then outlined in Section 3. The non-parametric hand model is described in Section 4, while Section 5 describes how inference is done over this model. Experiments in Section 6 show the non-parametric method to outperform other state of the art regression methods. We also show qualitative reconstruction results for a number of synthetic and real test sequences.

2. Related Work

In this section we review related work on object-contextual non-parametric hand pose estimation. For a general review on human motion estimation we refer the reader to [2] and for hand pose estimation in specific to [16]. Further, we will discuss the main difference, both with respect to accuracy and performance, of generative and discriminative methods in the context of hand pose estimation.

2.1. Object-Contextual Hand Pose Estimation

As discussed in the introduction, hand pose estimation can be addressed in a generative or a discriminative manner. Object-contextual hand pose estimation has been addressed in a generative manner in two recent works. In [12] the authors show how the hand pose can be reconstructed robustly despite the object occlusion. The hand is observed using RGB-D imagery (with both range and color). To achieve robustness to partial occlusion of the hand from objects, the hand is modeled as a Markov random field connecting segments corresponding to the different bones of the hand skeleton. In this way, the non-occluded segments can guide the pose estimation of the occluded ones.

In [13], this is taken one step further, with explicit reconstruction of the object in 3D. By enforcing physical constraints on the hand pose from the object 3D surface and vice versa, the two pose estimation processes guide each other. A multi-camera system is used to estimate both the pose of the hand and the object with framerates between 0.5 and 2 Hz.

2.2. Generative and Discriminative pose estimation

As outlined in the introduction inference of hand pose from images have either been done using generative or discriminative methods. In contrast to [12, 13], we take a discriminative approach to object-contextual hand pose estimation. Over the next paragraphs we outline and discuss the main difference between generative model-based estimation methods and discriminative regression estimation methods to motivate our line of approach.

Accuracy. An important advantage of generative approaches is their (potential) accuracy, which is only limited by the precision of the hand model and the computational time available. In contrast, the accuracy of our discriminative non-parametric approach is fundamentally limited by the design of the database; it is not computationally tractable, using any approximation, to add enough new samples to the database in order to reach the accuracy of a generative tracker.

Initialization and error recovery. However, one disadvantage of generative models is their inherent local character. In most cases, the posterior distribution over the state space is highly multi-modal. The estimation procedure must therefore have a good prior state estimate. This can represent a problem in the initialization of the method. The tracking procedures in [12] and [13] were manually estimated.

Another inherent problem of locality with generative models is the recovery from errors; when the pose of a frame is wrongly estimated, subsequent frames will try to adapt such erroneous estimation to new frames. Since the temporal propagation model by nature is local, the method will then lose track.

Discriminative, detection-based methods are inherently global, because a new search independent of the previous ones is executed in every frame. In our system we encourage locality by using a temporal consistency model, see Section 5.2. However, since the likelihood in our model is sampled globally, hypotheses from new parts of the pose space are continuously picked up, ensuring that the tracker can recover from errors easily.

The locality of model-based solutions can be specially problematic for hand pose estimation because hand movements in real sequences can be very fast ($5m/s$ translational and 300 deg $/s$ rotational speed of the wrist [16]), breaking the locality assumption.

Computational efficiency. The joint estimation of hand and object pose in [13] presents another problem: computational load. The results shown with real sequences use eight cameras and the estimation time is 2 seconds per frame. Decreasing the number of cameras (and therefore the quality) can speed-up the system up to 3 Hz. The method of [12] requires 6 seconds per frame.

In contrast, our discriminative method runs in real-time, implemented in C++ on standard hardware.

2.3. Non-Parametric Hand Pose Estimation

Other hand pose estimation systems have used databases of hand views in a non-parametric manner [7, 8, 11, 17]. As discussed in the introduction, none of the three previously mentioned systems mentioned how to handle or take advantage from occlusions, and the experiments showed hands moving freely without any object occlusion. The main difference between our system and previous approaches is that we exploit contextual information, such as objects to estimate the pose of the hand.

In [11], the application of a specially designed glove circumvents several problems associated with hand-pose estimation, making the problem as well as the ap-

proaches significantly different. An evolution of that system can be found in [17], where the authors track the hands without the need of gloves. However, they can only track a very limited range of hand poses and movements.

The system described in [7] performs classification of human hand poses against a database of 26 basic shapes. This is adequate for their intended application, automatic sign language recognition. In contrast, our method aims to perform continuous hand pose estimation rather than isolated single-frame pose classification, which means that we can exploit temporal smoothness constraints to disambiguate the estimation.

The work from [8] can be regarded as the most similar to our work. However, like the two other approaches, they only take freely moving hands into regard.

3. Probabilistic Framework

We begin by explaining the notation used throughout the paper. At a specific time instant t, let $\mathbf{x}_t$ be the articulated hand pose and $\mathbf{y}_t$ the corresponding image observation.

Given a specific image observation $\mathbf{y}_t$, we wish to recover the associated pose parameters $\mathbf{x}_t$ generating the visual evidence. Formally we will refer to the relationship between the pose and the image space as the *generative mapping* f,

$$\mathbf{y}_t = f(\mathbf{x}_t). \tag{1}$$

A discriminative approach to infer the pose from an image is to model the inverse mapping f^{-1} as a function, using a regression model as in [18]. In a probabilistic formulation, this function estimates the likelihood density $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$.

In the case of hand pose estimation, this is known to be a highly ill-conditioned problem, since the image features are ambiguous; the same image observation $\mathbf{y}$ might origin from a wide range of different poses $\mathbf{x}$, making the likelihood density multimodal [19]. In order to proceed, several different approaches have been suggested: generative models [20, 12, 13] which directly model f, approaches which rely on multiple views [9], or methods that exploit the temporal continuity in pose over time [20, 21].

In this paper, our objective is a highly efficient method for situations where model-based generative approaches are inapplicable due to their computational complexity. Further, multiple views are not available in most applications.[2] We

[2]It should be noted that it is straight-forward in the present approach to employ image evidence

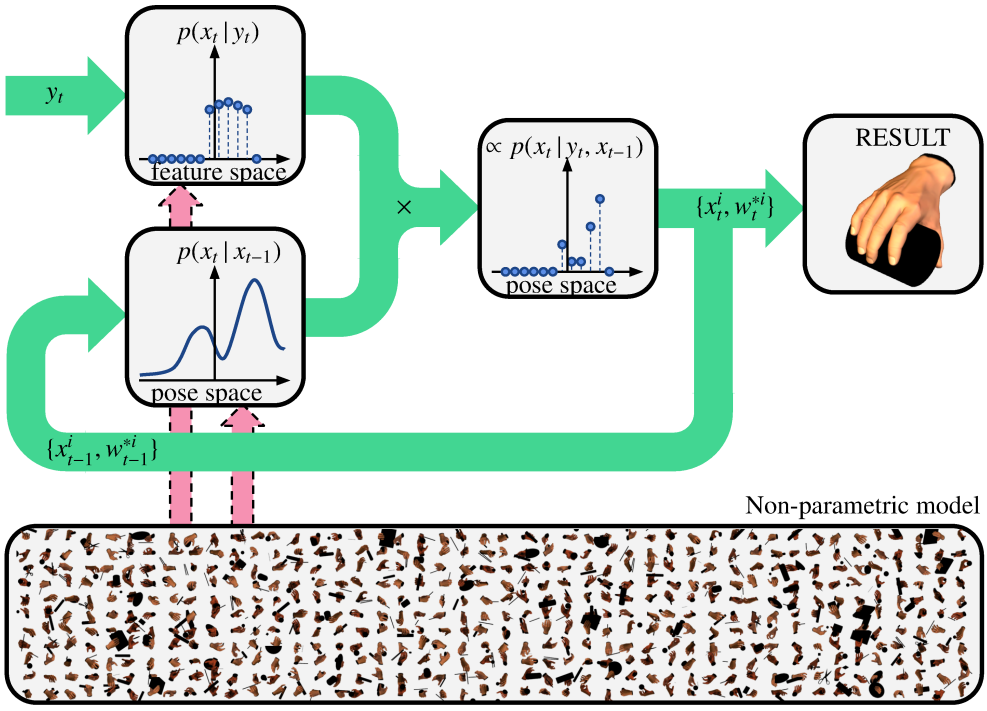


Figure 2: Schematic figure of the non-parametric temporal pose estimation framework. Given an image observation $\mathbf{y}_t$, a set of pose hypotheses $\mathbf{X}_t$ are drawn from the model. Each hypothesis is given a temporal likelihood based on consistency with the hypothesis in the previous frame. The final estimate is the pose associated with the largest probability.

thus take the latter approach and exploit temporal continuity to disambiguate the pose. The pose space is assumed to be Markovian of order one, i.e., the pose $\mathbf{x}_t$ depends only on the pose at the previous time step $\mathbf{x}_{t-1}$. The estimation task thus reduces to maximizing $p(\mathbf{x}_t \mid \mathbf{y}_t, \mathbf{x}_{t-1})$, which we decompose into an observation and a temporal model,

$$p(\mathbf{x}_t \mid \mathbf{y}_t, \mathbf{x}_{t-1}) \propto p(\mathbf{x}_t \mid \mathbf{y}_t) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \,. \tag{2}$$

In this paper we take a non-parametric approach, with an implicit likelihood model represented by a large database of images and their corresponding poses,

from several camera views, or alternatively from RGB-D imagery. This is also discussed in the Conclusions.

see Figure 2. To perform inference, we use a truncated approach where we approximate the distributions in Equation (2) using local models.

As shown in Figure 2, one time-step of inference is carried out as follows:

- Given an image observation $\mathbf{y}_t$, a set of weighted pose hypotheses $\mathbf{X}_t = \{\mathbf{x}_t^i, \mathbf{w}_t^i\}$ are drawn from the model as the nearest neighbors to the image observation in feature space. These constitute a sampled approximation of the observation likelihood $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$. This is described in further detail in Section 5.1.

- From the weighted nearest neighbors of the previous time step, a function $g(\mathbf{x}_t)$ approximating the temporal model $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ is computed. This is described in further detail in Section 5.2.

- Weights $\mathbf{w}_t^{*i}$ are now computed as $\mathbf{w}_t^{*i} = g(\mathbf{x}_t^i) * \mathbf{w}_t^i$. The weights are normalized to sum to 1 for all samples in $\mathbf{X}_t$.

- The pose estimate is the most probable sample from the database given the observation and the previous estimates. With our weighted nearest neighbor approach, this is approximated by $\hat{\mathbf{x}}_t = \mathbf{x}_t^k$, where $k = \arg\max_i \mathbf{w}_t^{*i}$.

In the next section we describe how the proposed implicit database model is created and represented.

4. Non-parametric Model Representation

In order to obtain the non-parametric model, we need to acquire a training data set of poses and associated image appearances $(\mathbf{x}, \mathbf{y})$ that can be assumed to "well" represent the problem, i.e., that includes poses that are expected to occur in a specific application domain. As our approach is non-parametric, there is no explicit parametrization of the image-to-pose mapping, as the relationship is implicitly parametrized by the database itself.

Generating such a database of natural images poses a formidable challenge, as it would need to capture the variations in pose and image appearance at a sufficient resolution in order to make accurate pose estimation possible. However, with recent advances in Computer Graphics we can use a rendering software such as Poser, which is capable of generating high-quality images of hands efficiently. The idea of acquiring large sets of training data using this approach is not new and has proved to be very successful for pose estimation [18, 4].

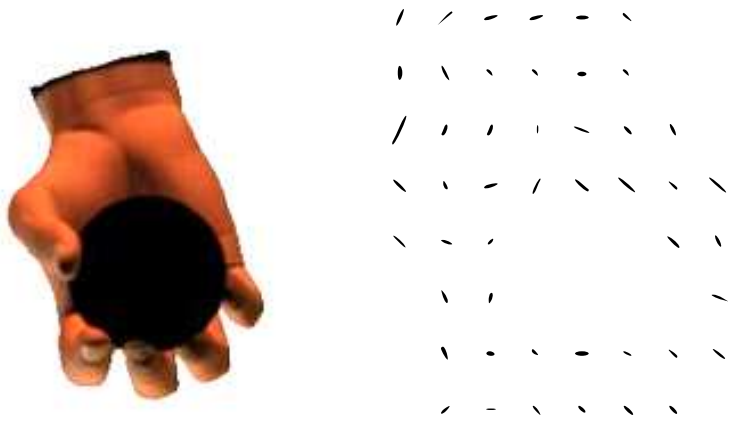


Figure 3: The left image shows an example from the database. The right image shows the associated image feature descriptor $\mathbf{y}$. Prior to extracting the feature descriptor the object is segmented from the image, resulting in a "hole" at the corresponding position in the descriptor. This encodes the correlation between pose and object in a more robust manner compared to if the internal edges of the object would also contribute to the descriptor.

The composition of the database used in this paper is motivated by our research aim: understanding human interaction with objects, [22, 23, 14]. We select 33 different grasping actions according to the taxonomy presented in [15]. Further, each action is applied to a set of basic object shapes on which the grasp would naturally be applied. Each action is then discretized into 5 different time-steps. In order to make our approach view-independent we generate samples of each instance from 648 different view-points uniformly located on the view-sphere. This results in a database of over 100 000 instances (see, e.g., Figure 3, left), which we assume samples the problem domain well.

4.1. Data Collection

Images are extremely high-dimensional objects, making it infeasible both in terms of storage and modeling to use the original pixel representation. In this paper we therefore apply a two stage feature extraction approach with the aim to remove variance not related to pose from the image. In the first stage the hand is segmented from the image using skin color thresholding [24]; this also removes the object being grasped and the parts of the hand occluded by the object. Having extracted the hand from the image, the dimensionality is further reduced by representing the image as the response to a image feature.

A large amount of work within Computer Vision has been focused on developing different image features [25, 26, 27]. An ideal image feature should be robust to segmentation errors, sensitive to non-textured regions and fast to compute. We compare the performance of Histogram of Oriented Gradients (HOG) [28] features and features based on distance transform [29] for different parameter settings. For a number of different feature options, the following experiment is performed: The feature is computed for every database entry. The entries are removed from the database one at a time, and the 50 nearest neighbors (NN) extracted from the database. The mean is taken of the Euclidean distance in pose space between all query entries and their found nearest neighbor number 1, 2, ..., 50. This distance is the same as the error of a non-parametric pose estimation – a dense database and a good feature would give small distances, while a sparse database and a non-informative feature would give large distances. Figure 4 shows the cumulative mean pose error of nearest neighbor number 1–50, for 9 different feature alternatives.

Based on the result shown in Figure 4, an $8 \times 8 \times 8$ HOG feature is selected, resulting in a 512 dimensional image representation, see Figure 3, right.

Our motivation is to exploit contextual information of the grasped object when estimating the hand pose; the object contains a significant amount of information about the pose (and vice versa). In a learning based framework, which assumes having a training data set which describes the problem domain well, the natural inclination is that the model would be limited to handle objects which are included in the database. Such a model would have to be of a size that would render it infeasible to use. However, in our model the object is removed. This means the occluding shape of the object affects the representation while the internal edges of the object do not, see Figure 3. This representation can robustly be extracted from the image and is capable of generalizing over different objects. As we will show in the experimental section, this sufficiently models the correlation between hand and object allowing estimation in scenarios of severe occlusion.

Having acquired a low-dimensional efficient representation $\mathbf{y}$ of the image as described above, the database is completed by associating each image $\mathbf{y}_i$ with its corresponding pose parameters $\mathbf{x}_i$. The pose vector $\mathbf{x}$ is composed of the rotation matrix of the wrist w.r.t. the camera and the sines of the joint angles of the hand.

5. Inference

As shown in Equation (2), the conditional probability density over hand pose $\mathbf{x}_t$ is factorized into two different terms, an observation likelihood $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$ and a

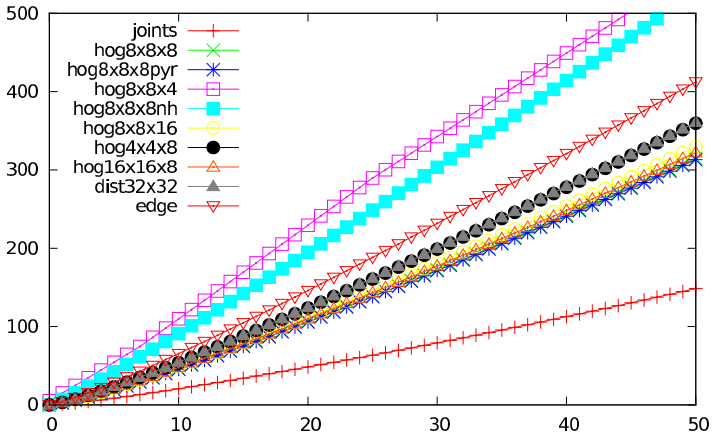


Figure 4: Cumulative mean pose error of non-parametric pose estimation using different image features. The curves show the cumulative Euclidean distance between the query pose and its nearest neighbor number 1-50 in the database. `joints` is the ground truth error in pose space, acquired by taking the nearest neighbors in the pose space directly. This is a lower bound on the error and shows the density of our database. The curves `hogAxAxB` show the error when using HOGs with $A \times A$ non-overlapping cells and a histogram of B bins (see Figure 3 for an example of an $8 \times 8 \times 8$ HOG). The suffix `pyr` indicates that the HOG feature includes lower resolution cells ($1 \times 1, 2 \times 2, \ldots, A \times A$). The suffix `nh` means normalized holes: the histogram is normalized to sum to one (i.e., removing information on how large part of the cell is covered by skin colored areas). The curve `dist32x32` shows the error when images are represented by their distance transform subsampled to 32×32 pixels. The `edge` curve shows the error when using the chamfer distance between edge maps extracted from the images. The result indicates that an $8 \times 8 \times 8$ HOG gives the lowest error.

temporal consistency model $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$. Below we discuss these two models in more detail, and show how the pose $\mathbf{x}_t$ is estimated from the observation $\mathbf{y}_t$ using the implicit database model.

5.1. Observation

The pdf $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$ is approximated by indexing into the database of hand poses using the image representation $\mathbf{y}_t$, and retrieving the nearest neighbors in the space spanned by the set of database features $\mathbf{Y}$. Due to the size of the database, an exact NN approach would be too computationally intensive. We therefore consider approximate methods. We compare Locality Sensitive Hashing (LSH) [30] and Fast

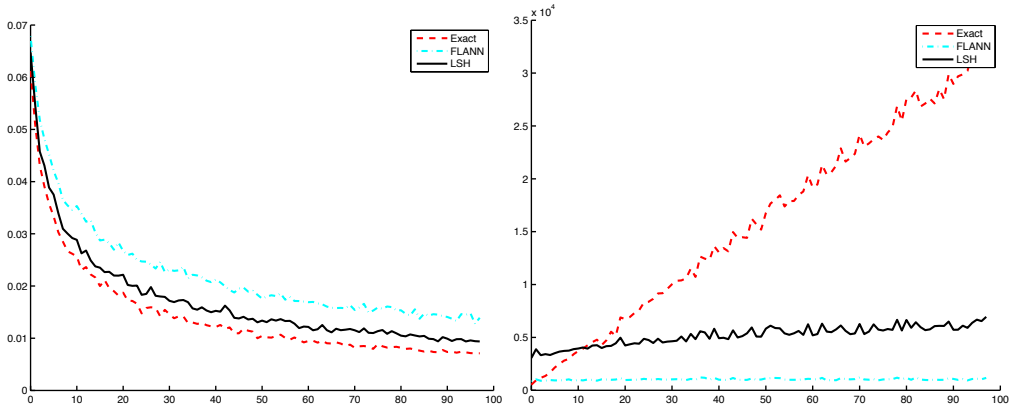


Figure 5: The plot shows the prediction error (left) and average query time (right) as a function of database size (as percentage of the full database size) for finding the nearest neighbor in the database. 10% of the original database is set aside for testing, resulting in a full database of around 90 000 instances. Two approximate methods, LSH and FLANN, are compared with an exhaustive search as baseline. The left plot shows that LSH performs slightly better than FLANN in terms of accuracy. The right plot shows the query time increasing linearly for the exhaustive search while the approximate methods being sublinear, and FLANN being faster than LSH in absolute terms.

Library for Approximate Nearest Neighbors (FLANN) [31], see Figure 5, and decide to use LSH in our experiments as it shows an attractive trade-off between computational complexity and prediction accuracy.

LSH is an ϵNN technique. This means that a query $\mathbf{y}_t$ results in an approximation to the exact nearest neighbor within a distance not more than $(1+\epsilon)$ times larger than the exact nearest neighbor distance. Each retrieved ϵNN $\mathbf{y}_t^i$ is associated a weight w_t^i from a spherical Gaussian density,

$$w_t^i = \mathcal{N}(\mathbf{y}_t^i \,|\, \mathbf{y}_t, \sigma_y \mathbf{I}) \,, \tag{3}$$

with standard deviation σ_y is set by experimental evaluation. This encodes our belief that the image feature representation is locally smooth and reduces the effect of erroneous neighbors from the LSH algorithm.

Each image feature in the database, $\mathbf{y}^j$ is associated with a pose $\mathbf{x}^j$. The poses $\{\mathbf{x}_t^i\}$ corresponding to the ϵNN $\{\mathbf{y}_t^i\}$ can thus be retrieved. Together with the weights, they form the set $\{\mathbf{x}_t^i, w_t^i\}$ which is a sampled non-parametric approximation of $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$.

5.2. Temporal Consistency

As described in Section 3, the temporal consistency constraint $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ is modeled as a parametric function g. It is used as a conditional prior to reweight the sampled distribution $\{\mathbf{x}_t^i, w_t^i\}$ approximating $p(\mathbf{x}_t \,|\, \mathbf{y}_t)$.

We assume that our model is getting observations densely enough in time such that the trajectory with respect to both the pose and view spaces vary smoothly. The naïve modeling approach would thus be to penalize estimates by their deviation in pose space to the previous estimate $\hat{\mathbf{x}}_{t-1}$. This model implicitly assumes that the temporal likelihood distribution $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ is uni-modal. The uni-modality assumption can introduce unnecessary errors in the prediction since $\hat{\mathbf{x}}_{t-1}$ might not be the best candidate due to ambiguities (several poses can share a similar appearance) or estimation errors. A more sensible approach is to make use of all the hypotheses $\mathbf{X}_{t-1} = \{\mathbf{x}_{t-1}^i, \mathbf{w}_{t-1}^{*i}\}$ in the previous time instance and propagate them through time. We can do so by modeling the conditional distribution $p(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ using a kernel density estimation (KDE) approach [32], where the density is modeled as a mixture of Gaussian kernels centered in $\mathbf{x}_{t-1}^i$ and weighted by $\mathbf{w}_{t-1}^{*i}$. This enables propagation of a potentially multi-modal distribution in time, making the temporal model significantly more flexible and expressive, allowing us to represent temporary ambiguities, resolving them further ahead in time.

As we will show in Section 6, having a strong temporal model allows us to perform prediction in noisy scenarios where the image observations are uncertain.

6. Experiments

We perform three sets of experiments using the proposed method. First we compare our non-parametric approach to a baseline of other state-of-the-art regression algorithms. In order to make an evaluation in terms of a quantitative error this experiment is performed using synthetic data where the joint configuration is known. Synthetic data also allows us to control the amount of noise in the images. Both our method and the baseline methods are evaluated in terms of robustness towards noise in the image observations.

In the second set of experiments we evaluate our method in a qualitative manner on synthetic sequences with added image noise.

The third set of experiments is performed on challenging real-world sequences.

We would like to encourage the reviewer to look at the videos submitted as additional material; they clearly show the robustness and generality of our approach.

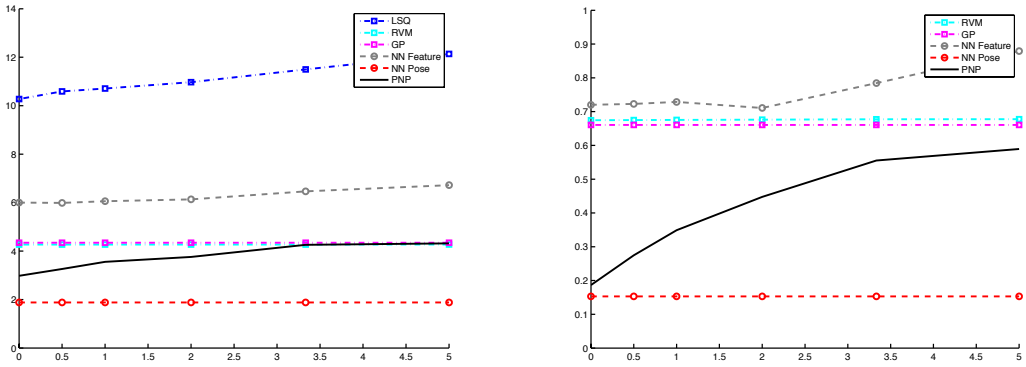
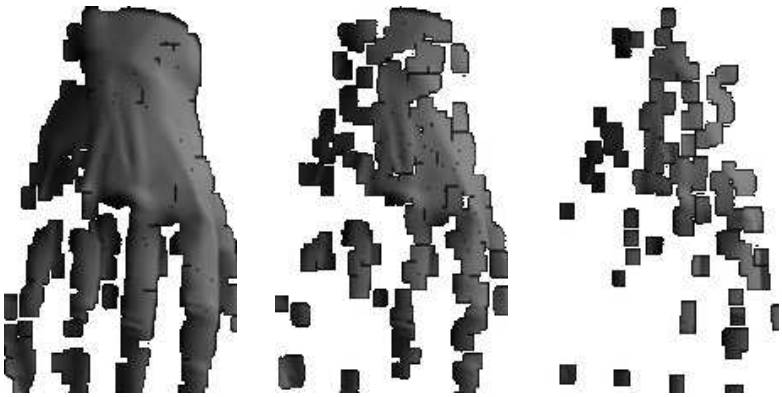
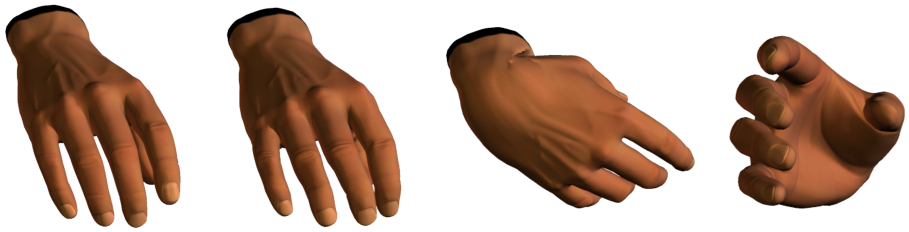


Figure 6: Pose estimation using the non-parametric method (`PNP`) in comparison to three different regression techniques (`LSQ`, `RVM`, `GP`). As a baseline, the true nearest neighbor pose error (`NN Pose`) is shown, as well as the pose error of the nearest neighbor in feature space, not taking temporal information into regard (`NN Feature`). The plots show the average error with increasing segmentation noise. The error measure in the left plot is the Euclidean distance in the pose space spanned by $\mathbf{x}$. The error measure in the right plot is the Euclidean distance in the space spanned by the 3D positions of all finger joints.



(a) $\alpha = 0.5\%$ (b) $\alpha = 3.3\%$ (c) $\alpha = 5\%$

Figure 7: Artificial corruption of the segmentation of the synthetic test data. The corruption is performed as follows: A partial segmentation is created by randomly removing α percentage of the pixels from the segmentation. The morphological operators of erosion and dilation are then applied this partial segmentation in order to propagate the noise over the image. Examples of increasing segmentation noise are shown.



(a) Original pose (b) Error 5.33° (c) Error 1.65° (d) Error 1.58°

Figure 8: Four different hand-poses are shown. The right-most image corresponds to the ground truth pose and the remaining images are estimates of the ground-truth. The estimates are ordered according to decreasing joint angle error. This clearly exemplifies how badly joint angle error corresponds to the quality of the estimate. This is because the norm in joint space assumes each dimension to contribute equally to the quality of the prediction. Therefore it does not reflect the hierarchical structure of the hand where error higher up in the chain (such as in the last two examples) effects the position of every joint further down the chain compared to the first prediction where the errors are concentrated closer to the finger tips.

6.1. Baseline

We compare our method to a set of regression models. In specific, we use Least Square Linear Regression (LSQ), the Relevance Vector Machine (RVM) [33] and Gaussian Process regression (GP) [34] to model the mapping from input features $\mathbf{y}$ to pose $\mathbf{x}$, approximating the likelihood $p(\mathbf{x}\,|\,\mathbf{y})$ (no temporal information is included here). Each of these models have previously, with significant success, been applied to pose estimation [18, 9, 35] for both hands and full body pose.

All above models are based on a fundamental assumption that the mapping f^{-1} from image to pose takes functional form; LSQ assumes linear form, while RVM and GP can model more flexible mappings. We compare these three methods to the suggested approach on four different synthetic sequences with varying degrees of added image noise, see Figure 7. Neither the poses nor the objects in the test sequences are present in the database.

As can be seen in Figure 6, left, the linear LSQ regression results in a very large error indicating that the relationship between feature and pose is strictly non-linear. The RVM and the GP are unable to model the mapping and do in fact always predict the same pose: the mean pose in the training data, irrespectable of image observation. In other words, this means that the appearance-to-pose

mapping f^{-1} is under-constrained and does not take functional form. However, the non-parametric approaches are capable to model in such scenarios. From the results we can see that an exact nearest neighbor estimate in the feature space (without temporal information) results in a worse result compared to the mean pose distance in the data set, while our approach performs significantly better – also indicating that the mapping is non-unique. The dashed red line shows the results of an exact nearest neighbor in the pose space and is therefore a lower bound on the error of our method as it shows the resolution of the database.

The norm in joint space is not easily interpretable in terms of quality of the prediction as it does not respect the hierarchical structure of the hand, see Figure 8. Therefore, the right plot of Figure 6 shows the same mapping results, but with an error norm in terms of finger joint 3D positions. This shows even clearer how well our suggested method performs. With very little noise we are close to the exact NN lower bound, with increasing segmentation error asymptotically moving towards the mean.

Note that 5% error corresponds to a very weak segmentation, see Figure 7. Further, our approach significantly outperforms the exact nearest neighbor in feature space (without temporal information). This clearly indicates how important temporal information is in order to disambiguate the pose.

To summarize, the results clearly show that the mapping from image features to pose is both highly non-linear and non-unique (multi-modal). This implies that it cannot be modeled using a functional approach.

6.2. Synthetic

In order to evaluate the qualitative performance of our method in a controlled scenario, we applied the model to image sequences with a controlled noise level. The results are visualised in Figure 9.

The estimated pose over the two sequences is accurate while the associated object varies. This validates our assumption that objects generalize over pose and provide important contextual information is correct.

6.3. Real Sequences

In order to show the performance of our method in a real world manipulation scenario, we let three different subjects, two men and one woman, manipulate three different objects. The objects are not contained within the model. The results are shown in Figure 10.

As can be seen from the results, our model is capable of accurately predicting the pose of the hand. In each of the sequences the test hand shape and appearance

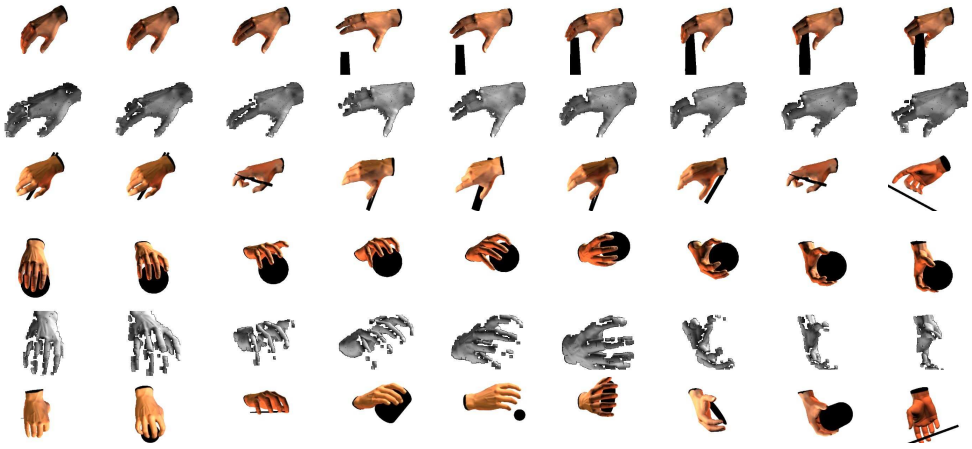


Figure 9: Qualitative results of our approach applied to synthetic data. The top and the forth row show the ground truth pose, the second and the fifth row show the segmentation from which the image features are computed. The segmentation has been corrupted by artificial noise with $\alpha = 0.5\%$ as explained in Figure 7. The third and last row show the corresponding predictions from our system. The two grasping sequences are applied to two different objects, in the first sequence a book and in the second a ball. We show the predicted hand-pose but also the object that is associated with the specific pose in the database.

is different from the database hand model, while there is no observable degradation in performance, showing that our model is robust to different hands. Further, as neither of the manipulated objects are represented in the model this further supports the notion that grasps generalize over objects and that the objects' influence on the grasp provide important cues. This clearly shows that our system is capable of exploiting such information.

A large portion of the dynamical models that have been proposed to the problem of pose estimation are based on auto-regressive models [36], which assumes that the trajectory in time takes functional form. Even though our dynamical model is parametric, it is based on hypotheses from the non-parametric ϵNN model. This means that it is considerably more flexible and can recover from bad estimates in situations where an auto-regressive model will fail. To highlight this strength we tested our model to a set of highly challenging sequences with fast non-linear motion and significant occlusion. This results in significant errors in the visual features. In Figure 11 the results clearly show the strength of our

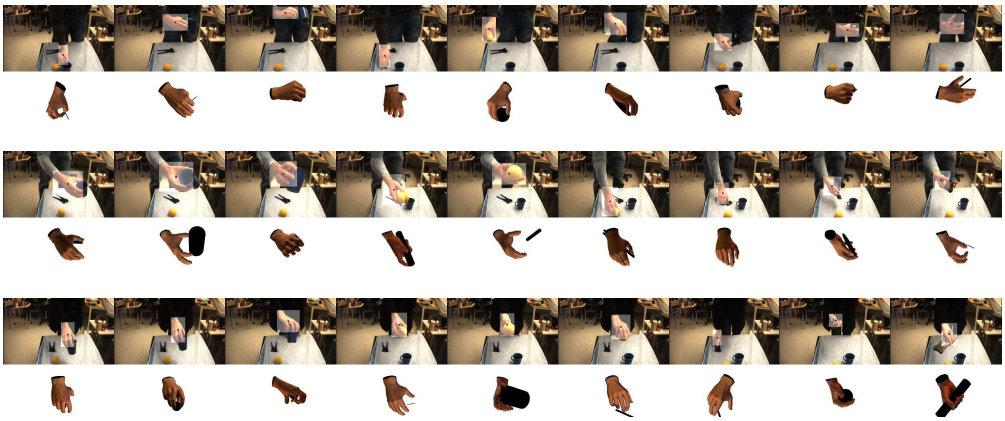


Figure 10: Predictions of real world sequences. The three rows show three different sequences where different objects are manipulated by different humans. In the first and second sequences the subject is male while in the last one female. None of the objects exist in the database. The first, third and fifth row show the input images with the skin detection window highlighted. The remaining rows show the associated predictions. As can be seen, the model correctly predicts the hand pose in each of the three different sequences.

approach, as it is able to track in such scenarios, and recover from errors which are difficult to avoid.

Further, we would like highlight the efficiency of our algorithm. It runs in real-time which makes it applicable in many different scenarios where pose estimation is an important source of information.

7. Conclusions

We present an efficient non-parametric framework for full 3D hand pose estimation. We show through extensive experimentation that the proposed model is capable of predicting the pose in highly challenging scenarios corrupted by significant noise or with rapid motions. Further, our model is efficient and runs in real-time on standard hardware.

The fundamental contribution is a system capable of exploiting contextual information in the scene from the interaction between the hand and a potential object. We show how this information can be exploited in a robust manner, making our system capable of generalizing the pose over different objects. This enables

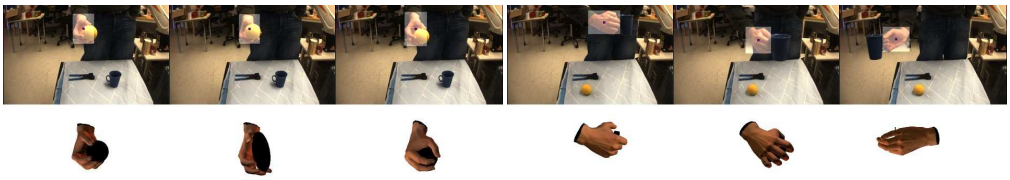


Figure 11: The above sequences shows two challenging examples. In the left sequence a significant portion of the hand is occluded by the object. However, our proposed method still manages to correctly estimate the pose of the hand. This clearly shows the strength of jointly estimating the object and the pose rather than seeing them as independent. The right sequence is an example where the subject manipulates the objects in a rapid fashion in a highly non-linear manner. In such scenarios most dynamical models commonly applied in pose estimation will over smooth the solution or be unable to predict at all due to being fundamentally auto-regressive approaches. Our model correctly predicts the pose in the two first frame while the last estimate is erroneous. This error is an implication of the Markov one assumption in our temporal model which thereby is not capable of modeling inertia and therefor is unable to resolve the ambiguity in the image sequence.

the usage of a fast discriminative method to scenarios where only expensive generative methods previously would have been applicable. We employ a multi-modal temporal model, allowing us to resolve ambiguities through temporal consistency. Our model could easily be extended to simultaneously estimate both the hand pose and the object shape by appending the inference scheme with a smoothness term with respect to object.

In future work we would like to evaluate the possibility of exploiting a better pose representation. This would make it possible to even further strengthen the temporal model. In this paper we also assume that the observation model can be modeled using a spherical Gaussian; this encodes an assumption of equal importance of the joint angles. This is unlikely to be true why we would like to explore a likelihood model that better respects the correlation between quality of estimate in joint space. This could potentially allow us to use additional hypotheses for each estimate.

Another avenue of future work to investigate is exploitation of RGB-D data, which would improve both the hand-background segmentation (currently based on skin color) and the feature representation of hand shape (currently HOG).

Finally, as noted in Section 2, generative and discriminative approaches have different merits. For applications requiring high accuracy, we therefore plan to run our discriminative hand pose estimator in parallel with a more accurate but less ro-

bust generative tracking method, using the discriminative estimates to (re)initialize the generative process.

References

[1] J. Romero, H. Kjellström, D. Kragic, Hands in action: real-time 3D reconstruction of hands in interaction with objects, in: IEEE International Conference on Robotics and Automation, 2010. 1

[2] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in computer vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2–3) (2006) 90–126. 2, 5

[3] A. Gupta, A. Kembhavi, L. S. Davis, Observing human-object interactions: Using spatial and functional compatibility for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1775–1789. 2

[4] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, in: IEEE International Conference on Computer Vision, 2003. 2, 9

[5] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011. 2

[6] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human-object interaction activities, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010. 2

[7] V. Athitsos, S. Sclaroff, Estimating 3D hand pose from a cluttered image, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003. 2, 6, 7

[8] B. D. R. Stenger, A. Thayananthan, P. H. S. Torr, R. Cipolla, Model-based hand tracking using a hierarchical bayesian filter, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1372–1384. 2, 6, 7

[9] T. E. de Campos, D. W. Murray, Regression-based hand pose estimation from multiple cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006. 2, 7, 16

[10] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr, R. Cipolla, Pose estimation and tracking using multivariate regression, Pattern Recognition Letters 29 (9) (2008) 1302–1310. 2

[11] R. Y. Wang, J. Popovic, Real-time hand-tracking with a color glove, ACM Transactions on Graphics 28 (3). 2, 6

[12] H. Hamer, K. Schindler, E. Koller-Meier, L. Van Gool, Tracking a hand manipulating an object, in: IEEE International Conference on Computer Vision, 2009. 2, 5, 6, 7

[13] I. Oikonomidis, N. Kyriazis, A. A. Argyros, Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints, in: IEEE International Conference on Computer Vision, 2011. 2, 4, 5, 6, 7

[14] H. Kjellström, J. Romero, D. Kragic, Visual object-action recognition: Inferring object affordances from human demonstration, Computer Vision and Image Understanding 115 (1) (2011) 81–90. 2, 10

[15] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, D. Kragic, A comprehensive grasp taxonomy, in: RSS Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, 2009. 4, 10

[16] A. Erol, G. N. Bebis, M. Nicolescu, R. D. Boyle, X. Twombly, Vision-based hand pose estimation: A review, Computer Vision and Image Understanding 108 (2007) 52–73. 5, 6

[17] R. Wang, S. Paris, J. Popovic, 6d hands: Markerless hand-tracking for computer aided design, in: ACM Symposium on User Interface Software and Technology, 2011. 6, 7

[18] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58. 7, 9, 16

[19] J. Romero, H. Kjellström, D. Kragic, Monocular real-time 3D articulated hand pose estimation, in: IEEE-RAS International Conference on Humanoid Robots, 2009. 7

[20] C. H. Ek, P. H. S. Torr, N. D. Lawrence, Gaussian process latent variable models for human pose estimation, in: Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2007. 7

[21] R. Urtasun, D. J. Fleet, P. Fua, 3D people tracking with gaussian process dynamical models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006. 7

[22] S. Ekvall, D. Kragic, Interactive grasp learning based on human demonstration, in: IEEE International Conference on Robotics and Automation, 2004. 10

[23] S. Ekvall, D. Kragic, Grasp recognition for programming by demonstration tasks, in: IEEE International Conference on Robotics and Automation, 2005. 10

[24] A. A. Argyros, M. I. A. Lourakis, Real time tracking of multiple skin-colored objects with a possibly moving camera, in: European Conference on Computer Vision, 2004. 10

[25] A. Kanaujia, C. Sminchisescu, D. N. Metaxas, Semi-supervised hierarchical models for 3d human pose reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007. 11

[26] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110. 11

[27] G. Mori, S. Belongie, J. Malik, Efficient shape matching using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1832–1837. 11

[28] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005. 11

[29] G. Borgefors, Distance transformations in digital images, Computer Vision, Graphics and Image Processing 34 (3) (1986) 344–371. 11

[30] W. Dong, Z. Wang, M. Charikar, K. Li, Efficiently matching sets of features with random histograms, in: ACM Multimedia, 2008. 12

[31] M. Muja, FLANN, fast library for approximate nearest neighbors, http://mloss.org/software/view/143/ (2009). 13

[32] V. Morariu, B. Srinivasan, V. Raykar, R. Duraiswami, L. Davis, Automatic online tuning for fast Gaussian summation, in: Neural Information Processing Systems, 2008. 14

[33] M. E. Tipping, Sparse Bayesian learning and the relevance vector machine, Journal on Machine Learning Research 1 (2001) 211–244. 16

[34] C. E. Rasmussen, Gaussian processes in machine learning, Advanced Lectures On Machine Learning: ML Summer Schools 2003 (Canberra, Australia, Tübingen, Germany). 16

[35] X. Zhao, H. Ning, Y. Liu, T. Huang, Discriminative estimation of 3D human pose using Gaussian processes, in: IAPR International Conference on Image Processing, 2008. 16

[36] J. Wang, D. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 283–298. 18

Visualization of Anthropomorphic Hand Performance

Thomas Feix, Javier Romero, Carl Henrik Ek, Heinz-Bodo Schmiedmayer, Danica Kragic

***Abstract*—The human hand serves as an inspiration for robotic and prosthetic hands. In the design of hand prostheses, an open question is which degrees of freedom to actuate in order to achieve the best functionality of the hand. In robotics, apart from the actuation, the goal is also to develop highly dexterous hands. A natural question is how to define a similarity measure through which the capabilities of different hands can be analyzed. Many parameters can be taken into account - ranging from kinematic and dynamic properties to the choice of material (rigid vs. soft) and interaction with objects. Currently, there are no analytic methods for performing such analysis and the mainstream approaches perform exhaustive experimental evaluation.**

In this paper, we address the problem of comparing the capabilities of different hands through the use of non-linear dimensionality reduction techniques. We concentrate on the kinematic analysis - that is, we address the problem of how many different grasp types or how large space of poses different kinematic structures can achieve. In our study, we first generate data with human subjects, thus using the capabilities of the human hand as the benchmark. The generated human data is based on an extensive grasp taxonomy, including most common grasp types. We develop a methodology for comparing different anthropomorphic robotic and prosthetic hands for the specific task of object grasping. We show how different robotic hands perform with respect to the human hand, resulting also in a comparison between different robotic hand designs. Although the method is applied to hand data, it can be used to compare other types of kinematic structures as well.

***Index Terms*—Grasping, Multifingered Hands, Kinematics, Rehabilitation Robotics, Biologically-Inspired Robots**

I. Introduction

We use our hands for daily interaction with the environment: the objects we interact with have been made to suit our dexterity. From robots we expect no less - they should be able to interact with and manipulate objects in the same/similar way we do. The same is desired for prosthetic hands. Historically, the road of building artificial hands has stretched between building simple industrial grippers and designing more complex hands that mimic human hand anthropomorphism and dexterity, [1].

In order to achieve the latter, one can add more actuators to a hand, resulting in a higher number of independently controlled joints. However, the effective dexterity of such a hand may not be increased due to the control complexity, [2]. Some of the mainstream approaches in robotics have the goal of creating hands that are relatively simple but still versatile in terms of the actions they can accomplish. The natural question arises: How sophisticated hands should we build, and how should we design them in order to be able to fully exploit their capabilities?

Several important works in robotic hand design have been inspired by the human hand, [3], [4], [5]. Relation between the human and prosthetic hands is clear given the similarity in the kinematic structure, but the problem of which degrees of freedom are best to actuate remains open. Thus, a natural question is how to define the metrics and perform analysis of the capabilities of different hands. Here, many parameters can be taken into account - ranging from kinematic and dynamic properties to the choice of material (rigid vs. soft) and interaction with objects. Currently, there are no analytic methods for performing such analysis. The mainstream approaches perform exhaustive experimental analysis and there is no unified benchmark for the problem.

In this paper, we address the problem of comparing the capabilities of different hands through the use of non-linear dimensionality reduction techniques. We concentrate on the kinematic analysis thus addressing the problem of how many different grasp types or how large space of poses different kinematic structures can achieve. In our study, we first generate data with human subjects, thus using the capabilities of the human hand as the benchmark. The generated human data is based on an extensive grasp taxonomy, including most common grasp types, thus with the focus on prehensile movements.

The main contribution of the paper is a methodology for comparing different anthropomorphic robotic and prosthetic hands. The methodology is based on the definition of an anthropomorphism index (AI) which measures the similarity between kinematic structures. The specific structures evaluated in our work are human and different artificial hands, both robotic and prosthetic. The approach allows for reasoning about the level of anthropomorphism of the artificial hand but can in general be used to assess the similarity between any types of kinematic structures. In our approach, the computation of the AI is based solely on a kinematic model of the hand and it is therefore straightforward to change parameters of the hand model and determine their impact on the AI. Parameters that can be changed are, for example, the number of joints, their orientation, etc. This provides a fast way of generating and assessing a changed hand design, providing the basis for its incremental improvement. An additional contribution of

This work was supported by EU IST-FP7-IP GRASP and Swedish Foundation for Strategic Research.

T. Feix is with Otto Bock Healthcare Gmbh, 1070 Vienna, Austria Thomas.Feix@ottobock.com

H.B. Schmiedmayer is with Institute of Mechanics and Mechatronics, University of Technology, 1040 Vienna, Austria heinz-bodo.schmiedmayer@tuwien.ac.at

J. Romero, C. H. Ek and D. Kragic are with Computational Vision and Active Perception Lab, Centre for Autonomous Systems, School of Computer Science and Communication, The Royal Instutute of Technology, SE-100 44 Stockholm, Sweden jrgn,chek,dani@kth.se

the work is the demonstration of how state of the art non-linear dimensionality reduction techniques can be used for this purpose, encoding the sparse high-dimensional data in a compact manner.

Finally, the benchmark procedure that is made publicly available through an open-source toolbox [1].

The paper is organized as follows. In Section II we overview the related work followed by the general idea proposed in our work in Section III. In Section IV we present the dimensionality reduction technique used and continue with the description of the metrics used for the comparison in Section V. In Section VI, we evaluate the latent space representation and in Section VII we present and evaluate two prosthetic and one robotic hand. Section VIII discusses the experimental results and concludes the paper.

II. Related Work

A human hand model typically consists of 20 independent joints, [6], [7], [8]. Studies have shown that this is an over-representation in terms of degrees of freedom (DoF) as there are strong correlations between the joints [9], [10], [11]. The correlations are not obvious and cannot be modeled explicitly, so data-driven approaches are commonly used to determine the coupling between them. The basic result is that only a few parameters are sufficient to unambiguously define a hand posture [9] or hand movement [12], [10], [11]. The minimum number of parameters required to specify the posture of the hand is called the intrinsic dimension of the hand or the number of DoF of a hand.

In robotics, a significant effort has been made at creating highly sophisticated hands with the goal of mimicking the versatility of the human hand. A few well known examples are the UB Hand 3 [13] with 16 DoF, the Robonaut Hand [4] with 12 DoF and the DLR-HIT Hand II with 15 DoF [5]. These hands have a large potential dexterity by design, but the real dexterity is much lower due to the control complexity, [2]. Apart from building simple industrial grippers, the recent focus has shifted from complex to simpler hands that can accomplish the assigned tasks [14].

The mechanical complexity of a hand and the complex hand-object interactions make it difficult to assess the quality of a hand design without its realization. Furthermore, there is no common benchmark for grasp performance measures. The classical way of determining the quality of a grasp is to assess the stability of a grasp and to determine whether a grasp is form or force closure [15]. The methodology is based on assessing how positional perturbations are resisted by the grasp, whereas in form and force closure the ability to resist external forces is determined. Such a measure can be used as a guideline for the hand design, as a hand should be built so that it has a good "stability score" on many different objects. A more elaborate comparison of different grasp similarity measures can be found in [16]. However, only a few hand prototypes are based on a structured analysis of their capabilities [16].

[1] http://grasp.xief.net

There are other relevant approaches to hand design optimization. One approach [17], [18] has been specialized towards underactuated kinematic hand setups. The actuation parameters of a hand are optimized to maximize the number of stable grasps achieved within a manually defined pool of grasping postures. The creation of the evaluation grasp set constitutes a time-consuming process. In [18], a prototype was built for a simple symmetric 2-finger gripper. For this special case it was possible to calculate a global optimal solution. For more complex embodiments, the objective function becomes more difficult to handle, having multiple local minima.

An approach using postural synergies is presented in [19]. A number of in-hand rigid-body object motions and internal forces can be applied to the object depending on the number of synergies (defined as basis vectors of a linear subspace) used to drive the hand. It was shown that with increasing number of synergies, more movements and forces become controllable. For example, having three contact points on the object and controlling the hand with one synergy, one internal force is controllable. Increasing the number of synergies up to three will render up to three internal forces controllable. A further increase in the number of synergies allows for intrinsic movements and finally, if the number is increased even more, redundant movements are possible. This analysis is a good tool to judge how complex a hand has to be in order to achieve a desired degree of dexterity. Nevertheless this tool is limited to linear subspace analysis and therefore to joint couplings in a linear combination sense. Nonlinear couplings or other complex joint coordination patterns cannot be modeled using this approach. Further, this does not produce hints or ideas on the kinematic design of a hand.

A framework to test underactuated hands is presented in [20], [16], where the ability of different kinematic setups to grasp a cylindrical object either by pinch or power grasp is assessed. The system determines the hand's ability to stably grasp moving cylindrical objects, as well as the grasp resistance to external forces. However, the system is used to evaluate a symmetrical gripper, where all axes are parallel. It is not clear how the system could be applied to more anthropomorphic hands, where the joint axes are not parallel. This is particularly important for the thumb, since the kinematic structure of the human thumb is very different to the rest of the fingers.

Using the tendon driven ACT hand, [21] investigates how tendon coordination patterns influence the positional precision of the hand. The decrease in precision is measured after tying various tendons to the same actuator. A small reduction in actuators is possible without large penalties on the fingertip precision. Controlling 20 tendons with only 16 motors results in an error of about 30%, which is still acceptable in most applications. However, when each finger is driven by only two actuators, the error is twice that of the fully actuated hand. Finally, the authors in [1] assess how different types of robotic hand components affect the trade-off between robustness to clutter and grasp stability.

For prosthetic hands, the problem is similar but the lack of a proper interface between the human and the prosthesis is the major bottleneck [22]. It is very difficult for the

technical system (i.e. the controller of the prosthetic hand) to determine the intended hand movement of the human. Although many methods for advanced prosthetic hand control have been proposed, none have proved to be sufficiently reliable for commercial applications [23]. Also, hand weight and reliability are key factors of user acceptance [24].

The work in teleoperation is also related to our work since it requires mapping of human controller movements to the robotic system with different hand/arm kinematics. Most of the examples rely on measuring the fingertip locations/joint angles and then estimating the joint angles of the robotic hand using inverse kinematics [25], [26], [27].

In summary, determining the quality of a hand design and assessing the effect of different parameters on the resulting functionality is difficult and remains an open problem. The work presented in this paper provides a methodology for comparison as well as a publicly available benchmark data. In the following section, we provide the details of the developed methodology.

III. System Description and Methodology

The goal of our work is to, as it will be discussed in more detail later, assess the level of anthropomorphism of different five-fingered hands. When it comes to generating and representing hand actions or hand configurations, we can specify them by using a joint angle representation, or using fingertip poses. In this work, we will generate a number of different hand configurations, both for the human and artificial hands, and represent them by fingertip poses.

Each fingertip has six DoF (three rotations, three translations) and we use rotation matrices to represent rotations. Therefore the vector representing one hand configuration has $5 \times 12 = 60$ elements. Further motivation for the choice of rotation parametrization is given in Section IV-B. Although the dimensionality of the representation *space* is rather high, the actual dimensionality of the fingertip *data* occupies only a small part of the representation space.

We define the term *action manifold* $\mathcal{A}$, that represents all the postures or a chosen subset of postures a hand can reach. For example, we may generate an action manifold that represents all three-fingered grasps for a hand. The goal in this paper is to use the action manifold, represented in fingertip space, for evaluation of similarity between different hands. More specifically, we compare human hands to different robotic and prosthetic hands, focusing thus on five-fingered hands.

The dimensionality of the action manifold will depend on the type/capabilities of the hand: for example, the dimensionality for a simple gripper, which only allows for an opening and closing, is one. In our approach, hands with different action manifolds are represented and compared in the same fingertip space. In short, the dimensionality of the representation space is $\mathcal{A} \in \mathbb{R}^{60}$ and the dimensionality of the data and thus action manifold is $\dim(\mathcal{A}) \leq 60$.

A. Assessment using Action Manifolds

We present the basic idea of the approach in Figure 1. The figure shows a hypothetical visualization of the fingertip space and the idea of its use for assessing the similarity between three hands. In that space, the data that can be generated by a specific hand spans a certain volume. Let us assume

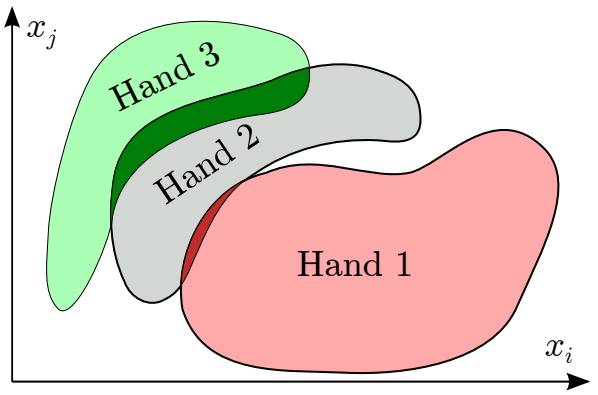


Fig. 1. Hypothetical visualization of the fingertip space $\mathbf{T}$ and of embedded action manifolds. Depending on the kinematics of the hand, the shape of the action manifold differs.

that "Hand 2" represents the volume populated by typical human motions while "Hand 1" and "Hand 3" represent the movements generated by two artificial hand setups. The aim of our approach is to estimate the intersection between the volumes spanned by two hands, *i.e.* to estimate which postures both hands are capable of generating. In the case of comparing a human and an artificial hand, the degree of overlap can reveal the level of anthropomorphism of the artificial hand. This overlap is denoted as its anthropomorphism index (AI) and it will be explained in more detail in Section V-D. We state again that the comparison is purely kinematic but if the artificial hand has soft fingers, for example, nothing in the methodology itself would need to be changed.

Comparing the volume of occupancy associated with the different hands is difficult as we do not know the actual density of the data but only have access to point estimates. In order to proceed we need to model the associated density of the data corresponding to each hand. This is a very ill-constrained problem, meaning that we need to make assumptions in order to proceed. Further, comparing such high dimensional data will be very expensive in computation terms due to the "curse of dimensionality". However, as we expect the dimensionality of the action manifold to be significantly lower compared to the fingertip space, we can exploit this when modeling the density. To that end we use a probabilistic dimensionality reduction approach which finds a parametrization of the density approximation using a single low-dimensional latent variable. The coordination of this variable will be the intrinsic parametrization of the action manifold.

B. System Overview

The first step is to generate an action manifold for a human reference hand which provides a basis for comparison with the prosthetic/robotic hand. To make the comparison and visualization feasible, the manifold spanned by the human hand motion is projected onto a lower dimensional space. This projection is performed using a dimensionality reduction

algorithm, described in Section IV-A. All possible fingertip configurations of an artificial hand are then projected onto that low-dimensional space. One example of this mapping is shown in Figure 2 as step 4. The white background represents all human hand movements projected to two dimensions and the colored trajectories are the projected movements of a prosthetic hand. We then compare how large is the overlap between these: a large overlap indicates that the hand is more similar to the human hand and thus more anthropomorphic.

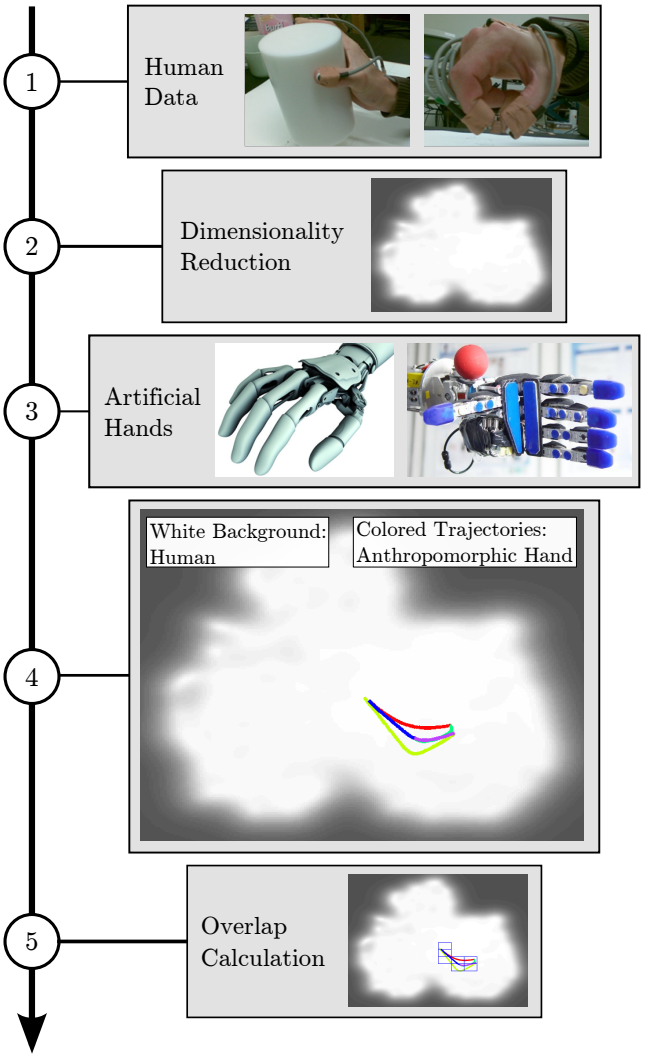


Fig. 2. System Overview: The recorded human hand movements (1) are projected onto a two dimensional space using a nonlinear dimensionality reduction algorithm (2). The white area represents all demonstrated human hand movements. The movements of an artificial hand (3) are then projected to that space (4) and the overlap is used as the basis for comparison (5).

The system consists of the following steps showed in Figure 2:

1) **Human data generation:** The first step is to generate a dataset of human grasping movements. These movements define the benchmark action manifold with which the manifold of artificial hand movements will be compared. Details on how this data was obtained are presented in Section V-A.
2) **Dimensionality reduction:** A nonlinear dimensionality reduction method is used to project the high-dimensional manifold to a lower dimensional space suitable for visualization and comparison. More details on the algorithm are given in Section IV-A.
3) **Artificial hand dataset:** Similar to the human dataset, a dataset of the movements of the artificial hand is generated based on its forward kinematics.
4) **Projection:** The artificial hand dataset is projected onto the low dimensional space spanned by the human data. The projection of artificial hands is done in Section VII.
5) **Overlap calculation:** The overlap between the manifolds is measured in the lower-dimensional space. In order to quantify the overlap, we created an overlap measure, the anthropomorphism index, which is explained in Section V-D.

We proceed by explaining the basis for the dimensionality reduction followed by the presentation of the data generation process.

IV. Dimensionality Reduction

As previously mentioned, the comparison of the Action Manifolds will be performed in a lower dimensional space. Consequently, we make use of state of the art non-linear dimensionality reduction techniques. The first reason for the choice of these techniques stems from the fact that the hand data is highly nonlinear. The second reason is that the techniques, as it will be discussed in the next section, provide the possibility of not only encoding the data in a fewer dimension but also of providing a likelihood measure. Finally, the employed technique gives us the possibility of encoding the high-dimensional data in a compact low-dimensional manner suitable for the comparison. In the next Section, we present the necessary details relevant for our work.

A. Gaussian Process Latent Variable Models (GP-LVM)

The Gaussian Process Latent Variable Model (GP-LVM) is a generative dimensionality reduction model. Let D denote the dimension of the data space and q the dimension of the low-dimensional latent space. Given N observations in the fingertip space $\mathbf{T}$, the matrix containing the data points is denoted $\mathbf{Y} \in \mathbb{R}^{N \times D}$ and the matrix of the corresponding points in the latent space is $\mathbf{X} \in \mathbb{R}^{N \times q}$. By assuming that the observed data has been generated through a functional mapping with additive Gaussian noise,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I})$, the likelihood $P(\mathbf{Y}|\mathbf{f})$ of the data can be formulated. The underlying idea of the model is to place a Gaussian Process ($\mathcal{GP}$)-prior over the generative mapping f. Combining this with the likelihood and integrating out the mapping leads to the marginal likelihood of the data,

$$P(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{j=1}^{D} \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} e^{-\frac{1}{2} y_j^T \mathbf{K}^{-1} y_j}, \tag{2}$$

where y_j is the j-th column of the data matrix $\mathbf{Y}$. The probability is calculated as the product of D independent

Gaussian Processes, each responsible for one dimension of the data space. The covariance matrix $\mathbf{K}$ defines the notion of similarity between points x_i, x_j and is constructed using a kernel function with the hyper-parameters θ. In this paper $\mathbf{K}$ takes the form of an RBF kernel combined with bias and white noise terms.

$$k(x_i, x_j) = e^{-\frac{\gamma}{2}(x_i - x_j)^T(x_i - x_j)} + \sigma_b + \sigma_n \delta_{ij} \tag{3}$$

Finally the solution to the latent locations and the hyper-parameters of $\mathbf{K}$ can be found by iteratively maximizing Eq. (2).

Back Constraints: In its basic form the GP-LVM does not guarantee the existence of a smooth inverse to the generative mapping [28]. However, this can be incorporated into the model by representing the latent locations x_i in terms of a smooth parametric mapping g_j from the observed data y_i;

$$x_{ij} = g_j(y_i, a) = \sum_{n=1}^{N} a_{jn} k_{bc}(y_i, y_n) \tag{4}$$

where k_{bc} is the back constraint kernel. This implies that the maximum likelihood solution of the parameters a rather than the latent locations are sought. This is referred to as a back-constrained GP-LVM [28]. In addition to constraining the latent location to preserve the local smoothness of the observed data, previously unseen data can be projected onto the latent space in an efficient manner by pushing them through this back-mapping.

We use a RBF (Radial Basis Functions) kernel of the following form:

$$k(y_i, y_j) = e^{-\frac{\gamma}{2}(y_i - y_j)^T(y_i - y_j)} \tag{5}$$

where the inverse kernel width γ controls the smoothness of the function. When projecting previously unseen points to the latent space, a sum over all contributions of the points from the training data is calculated.

B. Rotation Representation

The dimension of the original data space is dependent on the representation of the orientations. The data in our case involves three-dimensional position and orientation of the fingertips. This data will be interpreted as high dimensional vectors and compared in an Euclidean way both by PCA and GP-LVM. While the representation of positional data is straightforward, a representation of orientation which is "Euclidean-friendly" is less obvious. We explore different ways of representing orientation in the remaining of this subsection.

Euler angles are the most compact description of rotation in 3D space employing only three parameters. The big drawback of this method is the fact that the description with those 3 angles is not necessarily smooth even when the object moves smoothly in space. There are jumps in the data and additionally the method encounters the problem of singularities at certain rotations angles (gimbal lock) [29]. In other words, the result of a small change in orientation might be a big change in those three angles. Therefore, comparing the euler angles as three-dimensional vectors do not reflect properly changes in orientation.

Quaternions use four parameters to define the orientation. Three parameters can be interpreted as a vector and the last parameter is the rotation about this vector. Besides some computational advantages, this method is still very compact, and it offers smooth transitions from one orientation to the other without singularities [29]. The main drawback of quaternions is that the Euclidean distance between them does not reflect their similarity. Due to their properties the signs of the components of the quaternion can be inverted without affecting the transformation matrix [30, p. 162]. Therefore, the quaternion $\mathbf{q} = (e_0, e_1, e_2, e_3)$ represents the same rotation as $\mathbf{q}' = -\mathbf{q} = (-e_0, -e_1, -e_2, -e_3)$. The Euclidean distance between such a pair of quaternions is $\|\mathbf{q} - \mathbf{q}'\| = \|2\mathbf{q}\| = 2$, as quaternions are normalized $\|\mathbf{q}\| = 1$.

Rotation matrices use a 3×3 matrix which uniquely defines the orientation of an object at the cost of introducing additional dimensions. The rows of the rotation matrix can be seen as points whose position vectors correspond to an axis of the rotated system (Figure 3). This means that the Euclidean distance of their displacement varies smoothly with that of the orientation implying that the representation encodes the similarity we seek. By concatenating the matrix as a 9×1 point the Euclidean norm of the corresponding position vector will encode the joint displacement between the orientations as can be seen in Eq. 6,

$$R = \begin{pmatrix} x_0 & y_0 & z_0 \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{pmatrix}, R' = \begin{pmatrix} x'_0 & y'_0 & z'_0 \\ x'_1 & y'_1 & z'_1 \\ x'_2 & y'_2 & z'_2 \end{pmatrix}$$

$$\begin{aligned} \|R - R'\| &= \sqrt{\sum_{i=0}^{2}(x_i - x'_i)^2 + \sum_{i=0}^{2}(y_i - y'_i)^2 + \sum_{i=0}^{2}(z_i - z'_i)^2} \\ &= \sqrt{d_x^2 + d_y^2 + d_z^2}. \end{aligned} \tag{6}$$

C. Pose Representation

Using an Euclidean norm applied in the space of rotation matrices encodes similarity with respect to changes in orientation in a smooth manner. Clearly, as the fingertip location directly encodes a position, Euclidean distance will encode a sensible similarity between different locations. However, in order to apply the GP-LVM approach we need to compare different *poses* with each other and not independently rotations and positions.

To compare poses parametrized both as rotations and positions using a Euclidean norm we need to make sure that the relative scale of each dimension corresponds to similar "distortions" in terms of pose. To that end we independently transform the dimensions of the parameter space such that each possible configuration is contained within a hyper-cube.

This has the implication that we consider a translation of the length of a hand to correspond to rotation of π. Smaller rotations and translations are scaled accordingly. We believe that this encodes a sensible relationship between rotations and positions.

By bounding the parameter space to a hyper-cube we effectively encode an invariance to different hand-sizes. The

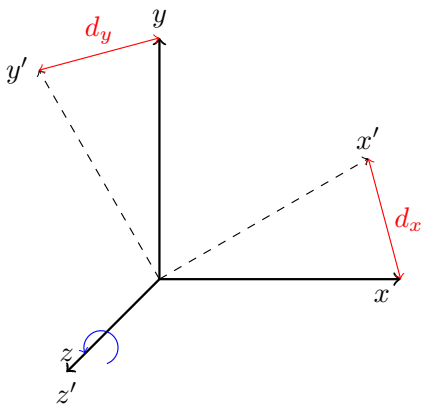


Fig. 3. Euclidean distance between rotation matrices R and R' interpreted as nine-dimensional vectors. An orientation R is rotated $30°$ around z axis to obtain R'. The distance $\|R-R'\|$ between those orientations is the Euclidean norm of a vector composed by the distances d_x, d_y, d_z between each of the axis normal vectors.

consequence of this is that our method does not take into consideration the absolute position of the finger-tips but only places relevance on the relative position between the different fingers. The motivation is that we are interested not in the kinematic capabilities of an artificial hand but only its similarity to the specific reference hand, normally a human hand.

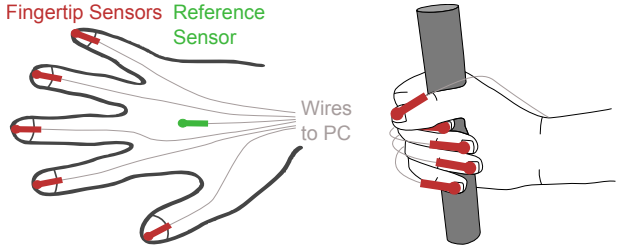
V. Data Generation and Encoding

In the following two sections we explain how the data was generated, both of the human subjects and for the artificial hands. This is followed by the section that explains how the data was encoded using the non-linear dimensionality reduction techniques in Section V-C. Finally, the estimation of the anthropomorphism index is described in Section V-D.

A. Human Action Manifold

In order to obtain a representative human action manifold, grasping data was recorded. The focus is one-handed static grasps. It is based on a measurement on five subjects (three male, two female); all subjects were right handed and did not report any hand disabilities. The average hand length and width were 185.2 mm and 81.1 mm respectively, with standard deviations 13.3 mm and 7.4 mm. A Polhemus Liberty system with six magnetic sensors was used for recording the data. The spatial and angular resolution of each sensor is 0.8 mm and 0.15 degrees respectively. A sensor was applied to the nail of each fingertip. An additional sensor was placed on the dorsum of the hand as a reference. See Figure 4(a) for how the markers were applied to the hand.

The subjects were asked to perform 31 different grasp types, as described in [31], on an object typical of each action. Initially, the hand was placed flat on the table next to the object to be grasped. Upon starting signal, the subject grasped the object with the desired grasp type, lifted the object (this moment is shown in Figure 4(b)), replaced it and retreated the hand to the starting position.



(a) Placement of the sensors. Five sensors are placed on the fingertips and one is positioned on the dorsum of the hand.

(b) Example grasp posture for grasp number 2.

The data recording started when the hand began to move and ended when the hand was returned to the initial position. Each grasp type was performed twice. The second trial is used for training and the first for testing (see Section VI-B). The fingertip sensors were transformed into the coordinate system of the reference sensor in order to remove global hand movement.

The resulting dataset consists of 4650 datapoints (30 samples $\times$ 31 grasp types $\times$ 5 subjects). Each fingertip described by a 12 dimensional vector (three encoding position, nine rotation).

By selecting a different reference dataset we can prioritize certain capabilities of the robot hand by deciding which actions/grasps are important. For example, if only very small objects are to be manipulated, there may be no need to include big object since it would promote hands with power grasp capabilities. As the focus in this paper is on general-purpose hands, the dataset is not restricted to a certain class of grasp types.

B. Robotic Action Manifolds

In this paper, the action manifolds of the robotic hands were obtained via kinematic hand models implemented in Matlab. The joint space of the hands is sampled and the corresponding fingertip configurations are determined. Furthermore the scaling of the dataset is in analogy to the human hand dataset - the positions are divided by the hand length and the orientations are transformed to rotation matrices.

There is no strict rule on how dense the sampling of the joint space has to be, but there is a guideline. The inter-point distance of the latent space projection should be smaller than the discretization of the latent space (the box length). Further increasing the density of the sampling will not increase the AI as the boxes are already populated by at least one point.

C. The Low Dimensional Space

The hand posture data is high dimensional and we expect it to be presented in a highly redundant parametrization. For this reason we use GP-LVM to embed the datapoints into a lower dimensional space which keeps the essential information about the hand posture. To create the low dimensional space we use the Matlab FGPLVM toolbox [32]. We chose a two dimensional space, as it is straightforward to visualize the comparison. In previous work, we have shown that the

relevant grasp information can be preserved in this lower dimensional space, [12]. The appearance of the latent space can be quite diverse, depending on the scaling of the data and parameters of the GP-LVM. By systematic variation of the GP-LVM parameters, 50 models were created and then evaluated according to the measures presented in Section VI. As will be presented in detail in that section, the models are compared in terms of their ability to distinguish between random hand models and the test set.

Regarding the human data set, the subjects generate a similar trajectory when performing the same grasp type which allows for a calculation of the mean grasp trajectory. To gain intuition on the structure of the space, Figure 4 presents five such trajectories. Depending on the grasp type, the final posture corresponds to a different location in the latent space. From the starting position on the right side, the subjects proceed to the final grasping point (indicated by circles) and then retreat back to the start position. There is a trend in the space that the further left a grasp type is located, the more the fingers will be flexed. This is natural since the starting posture is a flat hand and, as the fingers flex, the difference to that posture increases.

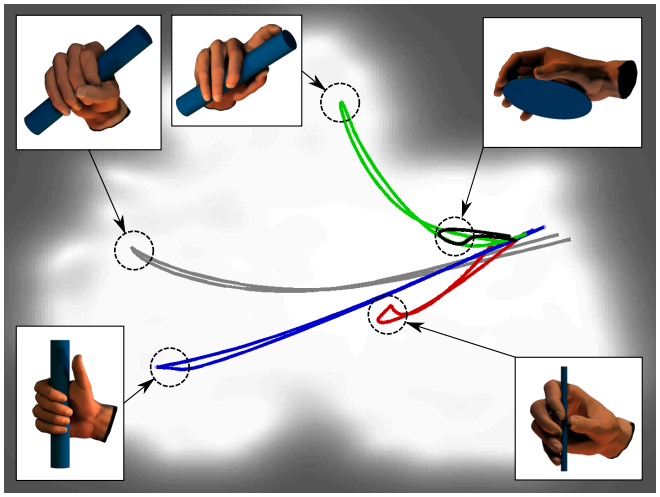


Fig. 4. Latent space representation of the grasping data. The trajectories correspond to the average trajectory of 5 subjects performing the shown grasps. Each grasp type is located at a distinct area in latent space. Only a few major grasp types are presented in the figure.

D. The Anthropomorphism Index (AI)

Once the movements of an artificial hand are projected onto the human spanned latent space, the overlap between that space and the human spanned manifold has to be measured. Our approach is to discretize the latent space into a regular grid and count how many cells are populated by a given hand design. For example, for a one DoF gripper the projection is a single line, whereas for a more complex hand with multiple actuators this can be a concentration of points with an arbitrary shape.

An important parameter for the calculation is the width of the cells as we regard all points within one cell as being equal. With equal we mean that if we vary the position within that margin, the resulting hand posture will only change by a small degree. As presented in Section V-A, each subject performed the grasp types twice. The difference in the final grasp posture of the hand of trial one and trial two can be regarded as being irrelevant as both configurations resulted in a stable grasp. Points belonging to the actual grasping poses of trial one and two are projected onto the latent space and the distance between two corresponding points is averaged over all trials and subjects. This gives a maximum distance d_x and d_y in x and y direction respectively which can be regarded as being the same grasp. Those lengths will define the resolution of the grid in latent space.

The GP-LVM models the mapping from the latent to high dimensional space using a Gaussian Process. This mapping provides us with a mean (prediction of the high dimensional location of the point) and a variance. The inverse of the variance, or confidence, is related to how certain the model is when reconstructing that point. The confidence C is scaled into the interval $[0, 1]$, where the white area in the latent space plots corresponds to maximal confidence. In regions where there are many data points, the variance of the projection will be very low. Consequently, the confidence will be close to 1. In sparse regions, the confidence will fall off as the projection gets more uncertain. A measure of the area of the human spanned latent space A_h can be calculated by summing the area of each cell $A_b = d_x \cdot d_y$ weighted by their corresponding confidence C_i.

$$A_h = \sum_i C_i \cdot A_b \tag{7}$$

The projection of the artificial hand movements discretized into M steps will result in a set of points $P \in \mathbb{R}^{M\times 2}$ whose overlap A_r will be calculated. This is done by summing over all cells which are populated by at least one point P_k.

$$A_r = \sum_i A_b \cdot \begin{cases} C_i & \exists P_k \in b_i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Finally A_r can be set into a relation to the area of human spanned space A_h and the relative latent overlap can be calculated. The ratio ${}^{A_r}/_{A_h}$ is multiplied by 100 to obtain a percentage value. We refer to this value as the anthropomorphism index (AI). It shows what percent of the human demonstration is covered by the robotic hand.

$$AI = \frac{A_r}{A_h} \cdot 100 \tag{9}$$

In the figures where the movements of artificial hands are projected to the latent space, we also plot the cells that were populated by the hand. That gives an idea on how the system works and additionally helps visualizing the overlap.

VI. Evaluation of the Latent Space

Our system looks for a latent space where the overlap between a human and a non anthropomorphic hand is minimal.

In GP-LVM, we can use different parameters to influence the structure of the latent space. One of these parameters is the inverse width of the back constraints kernel, γ. As described in Section IV-A, the projection from high to low dimensional

space is governed by back constraints. In Equation 5 we can observe that the ratio between the distance between points $y_i - y_n$ and the inverse width $\frac{1}{\gamma}$ determines the influence of different high-dimensional points y_n on the low dimensional point x_i.

Figure 5(a) represents a situation in which the kernel width is small compared to the inter-point distances. In this case, the support of any external point becomes negligible. New anthropomorphic data (white circles) will not be supported by our latent representation in this case, thus

$$\frac{1}{\gamma} \ll y_i - y_n \Rightarrow a_{jn} e^{-\frac{\gamma}{2}(y_i - y_n)^T (y_i - y_n)} \approx 0, n \neq i \quad (10)$$

On the other side, a large kernel width makes all points in the original space to equally support any point in the latent space (Figure 5(c)), no matter if they are anthropomorphic (white circles) or not (crosses), thus:

$$\frac{1}{\gamma} \gg y_i - y_n \Rightarrow a_{jn} e^{-\frac{\gamma}{2}(y_i - y_n)^T (y_i - y_n)} \approx 1, \forall n \quad (11)$$

Our goal is to use a value of γ such that only those points that correspond to anthropomorphic postures are taken into account, Figure 5(b).

As we cannot directly determine how well the kernel width represents the manifold, we ensure that the chosen width results in random models obtaining a very low AI, while human grasping data obtain a very high AI.

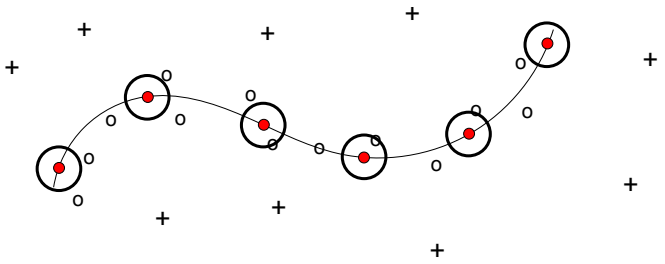
A. Random Models

We want the resulting space to have the ability to discriminate between human-like and non-human-like hands. Thus, when the movements of a non-anthropomorphic hand are projected onto the latent space spanned by the human hand movements, the AI should be minimal. This should occur even in the case of high dimensional non-anthropomorphic hands, which might have a large action manifold.

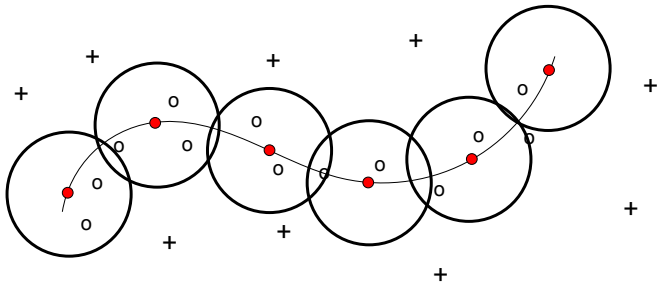
To test the behavior when projecting non-anthropomorphic hands, we create multiple random hand models. The random models are created using random Denavit-Hartenberg parameters with 3 DoF for each finger. Additionally, the positions and orientations of the bases of the fingers are random and the relative orientation of the fingertip coordinate frame to the kinematic chain is random. Joint angles are selected randomly from a 15-dimensional uniform distribution between 0 to 2π. Overall we take 20000 random samples from the joint space and calculate the corresponding fingertip poses. To model the hands we use a Matlab robotic toolbox [33] which allows us to calculate the fingertip poses through forward kinematics.

Figure 6(a) shows a typical representative of the set of random hand models. By simple inspection it becomes clear that this hand setup is not anthropomorphic. If we project the movements of four such random hands to the latent space (see Figure 6(b)) we see that all the movements collapse into a very limited region in the middle of the latent space. That demonstrates, that the AI of hands which are non-anthropomorphic is close to zero.

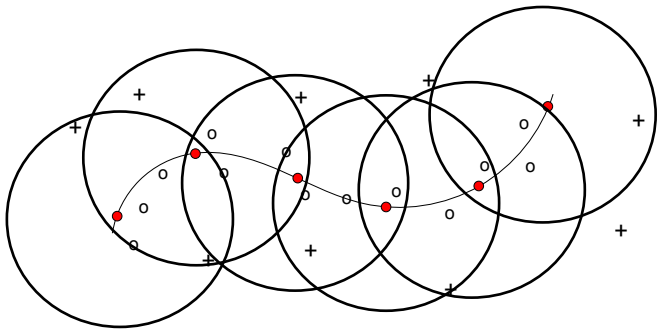
The model has the first desired property – hands that are different to the human cannot have a large AI score.



(a) The manifold is not represented properly as the kernel width is so small that there are places on the manifold where it is not supported by the training data.



(b) The manifold is represented correctly as the points of the test set can be supported by data whereas the more distant points of the random hand are not within the region supported by data.



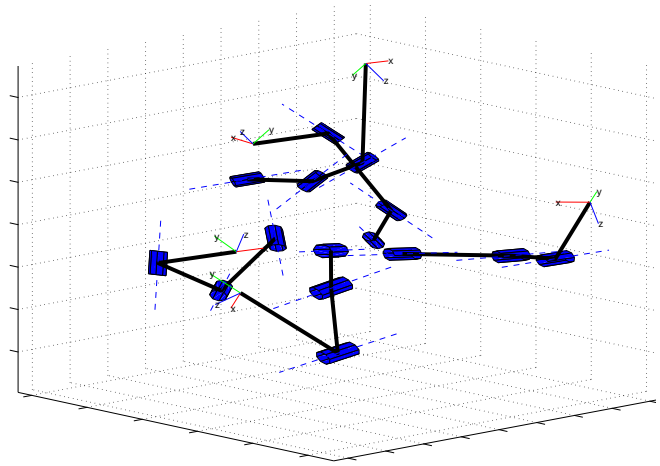
(c) The kernel width is so large that it will generalize over a too large part of the space.

Fig. 5. Different kernel width and their influence on the discrimination between the test set (empty dots) and the random hand set (crosses). The manifold (as indicated by the line) is sampled by datapoints from the training set (filled circles) and their corresponding kernel width (circles) is presented as well.

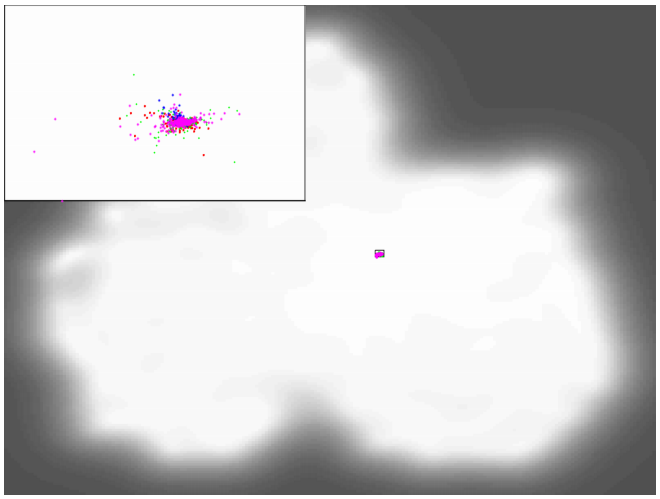
B. Test Set

The result of Section VI-A is that a non-anthropomorphic hand does not significantly overlap with the human spanned manifold even if it might have a large action manifold. In this section we will show that, given points which are similar to the training set, the whole latent space can be filled.

In order to verify this, we project the test set to the latent space. The test set is similar to the training set as in both cases the subjects succeeded in grasping the object with the demanded grasp type. The only differences between the sets is the variation introduced by the executions of the users. In Figure 7 the projection of the test set is shown. We observe that the test points are scattered accordingly to the training set. Yet, due to the width of the kernel there is a "halo" around the points which increases the area A_h and thus the AI score of the training set is reduced to 67%. This sets an upper bound for the maximal AI score artificial hands can achieve.



(a) Random Hand model



(b) Projection of four random hand models to the latent space. The left top picture is a magnification of the latent projection points. In that subfigure different colors indicate different hands.

Fig. 6. If random hands, like the one shown in (a) are projected onto the latent space, they cover only a very small area on the human manifold, as shown in (b). This means that the model is sensible enough so that a random model is not regarded as being anthropomorphic.

The previous two sections showed that the model is capable of distinguishing between hands that are human-like and those that are not. We can therefore proceed and use the system to benchmark existing prosthetic and robotic hands.

VII. Experimental evaluation

The anthropomorphism of two prosthetic and one robotic hand will be measured using the proposed methodology. A short discussion is provided for each hand, both regarding the data generation and comparison with the human hand.

A. SensorHand

The Otto Bock SensorHand [34] is a prosthetic hand (see Figure 8) with three actuated fingers which are all driven by

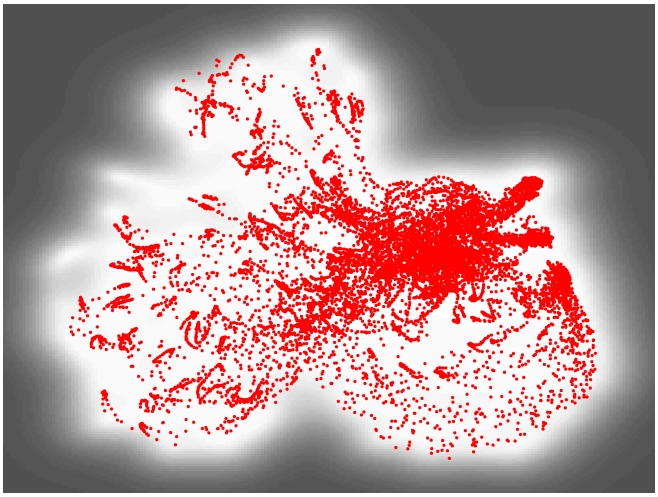


Fig. 7. Projection of the test set data to the latent space, the points cover most of the human spanned manifold (i.e. white area).

the same motor. The mechanical structure is covered by a glove, which is responsible for protecting the mechanics of the hand and creating a more human-like appearance. The glove also emulates the ring and the little finger, resulting in a 5-digit design. There is a metal bar within the glove which couples fingers four and five to the movements of the middle finger. As they are solely connected via the glove, the movement amplitude decreases from middle to little finger. The forward kinematics of the hand take that into account by reducing the maximal finger flexion of the ring and little finger. The finger angles α_i, where $i = 1$ is the thumb and $i = 5$ is the little finger, are depending on the driving variable a, where $a = 43°$ is hand closed and $a = 0°$ is hand opened. Overall 100 equally spaced samples of a were taken from that range. The corresponding finger flexion angles α_i are as follows.

$$\begin{aligned} \alpha_{1,2,3} &= a \\ \alpha_4 &= 0.9 \cdot a \\ \alpha_5 &= 0.8 \cdot a \end{aligned} \tag{12}$$

Fingertip poses are computed based on these flexion angles. Their projection during one opening-closing cycle, which is all the hand is capable of, is shown in Figure 9. The hands' AI is 0.25%. The trajectory is different to the projection of a random hand, see Figure 6(b).

Overall the hand has some major differences with the human hand. The position of the thumb is not anatomically correct; it is basically rotated 180 degrees, so that it perfectly opposes the index and middle finger. Even though the positions of the thumb fingertip are potentially correct, the orientations are not. The human cannot orient the fingertip in such a way as the SensorHand does. Additionally, all finger MCP[2] joints share the same rotation axis. A more natural way would be to orient the axes in such a way that the fingers are slightly abducted when the MCP joint is extended. All those non-anthropomorphic features combined are the reason why the latent space trajectory of the hand is relatively short.

[2]The Metacarpophalangeal joint is connecting the metacarpus to the first phalanges (fingers) in the human hand.

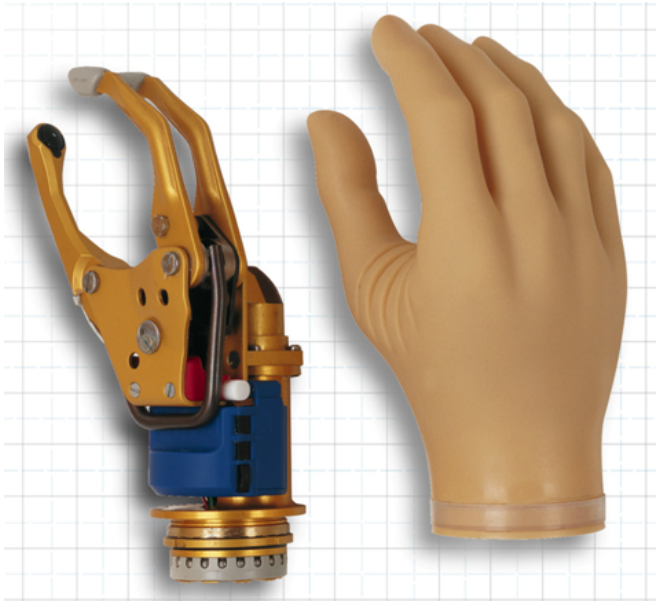


Fig. 8. The Otto Bock SensorHand: left) the hand without the covering glove; and right) the glove that is put over the hand for protection of the hand and for cosmetic reasons.

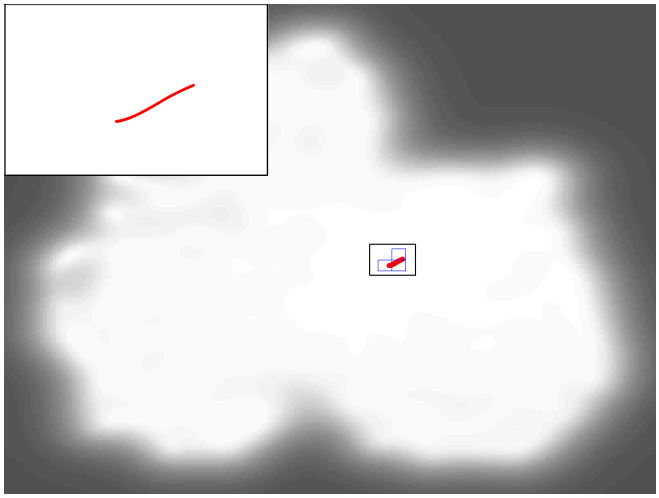


Fig. 9. Projection of the fingertip movements of the Otto Bock SensorHand to the latent space. The red points represent the trajectory of one open-close cycle. The hand has an anthropomorphism index of 0.25%.

As we use hand models, the properties of a hand setup can be changed and the effect on the latent space overlap can be analyzed. In our case we actuate independently the 5 joints which are coupled in the SensorHand (the CMC [3] of the thumb and the MCP joints of the fingers), conferring five DoF to the hand. The range of motion is the same as previously and we take 9 equally spaced flexion values for each joint. Overall this creates $9^5 = 59049$ different hand postures which are projected to the latent space. The projection (Figure 10) shows us that increasing the dimensionality of the hand does not change the latent space overlap much. The much more complex hand has an AI of only 0.4%,

[3]The Carpometacarpal joint is the most proximal joint of the human thumb.

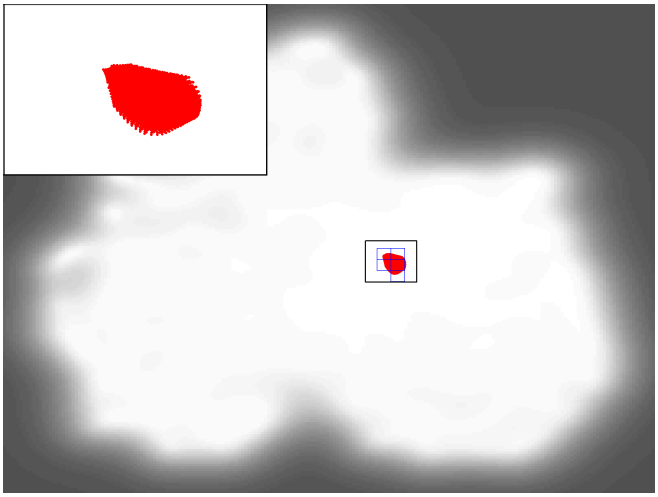


Fig. 10. Projection of the fingertip movements of the 5 DoF "Otto Bock SensorHand" onto the latent space. The hand has an AI of 0.4%.

which is a slight increase to the original SensorHand. Adding independent actuators proved to be a bad choice for increasing the hand anthropomorphism.

B. Michelangelo Hand

The next generation of prosthetic hands by Otto Bock is the so called Michelangelo hand [35], Figure 11. It follows a more human-like kinematic setup and it has 2 DoF. The axes of the finger MCP joints are oriented in a more natural way, where the flexion of the finger also invoke a small adduction. The fingers are slightly abducted when the MCP joints of the fingers are extended, whereas when flexed the fingertips touch each other. The first DoF is the main drive which is responsible for a coordinated flexion and extension of the five digits. The second DoF changes the thumb position – it can be abducted or adducted. As the hand is still in development, the exact control scheme of the hand is not yet finalized. Therefore, we used a current hand implementation that had the following hand postures:

- Hand open for tripod pinch (OT)
- Hand open for lateral pinch (OL)
- Neutral position (NP)
- Tripod pinch (TP)
- Lateral pinch (LP)

The following movement trajectories between positions are incorporated into the hand model.

- OT $\rightarrow$ TP
- OL $\rightarrow$ LP
- NP $\rightarrow$ OT
- NP $\rightarrow$ OL
- NP $\rightarrow$ TP

Each trajectory is sampled with 100 points and the corresponding fingertip poses are projected onto the latent space. Figure 12 shows the projection of those movements, where the colors indicate different trajectories. Compared to the SensorHand, it can be observed that the trajectories are much longer. Therefore they are able to achieve an AI of 2.8%. Even

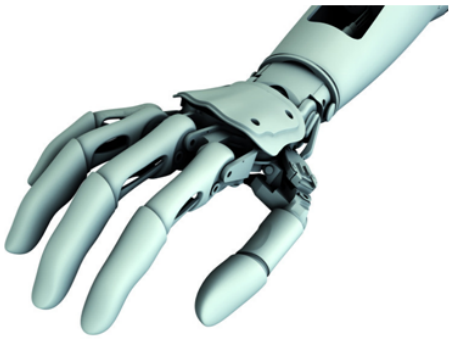


Fig. 11. Otto Bock Michelangelo hand [35].

though the hand still has very few DoF, its score is significantly larger as the general setup is closer to the human hand.

The tripod pinch (TP) and lateral pinch (LP) are located on the left side whereas the hand is opened on the right side of the latent space. In between lies the neutral position (NP) with trajectories connecting it to OT, OL and TP. If the movements of the Michelangelo hand are compared to the human grasp trajectories of Figure 4, it can be observed that they also show a left-right dominance and the starting position is on the right side whereas the grasp positions are on the left. That can be regarded as a sign that not only is the hand capable of covering larger areas in the human manifold, but also that the movements itself are human-like.

The positions of the tripod pinch and the lateral pinch are relatively close in latent space. That is due to the system roughly weighting every finger the same. As the poses of four of the five digits are nearly identical (in the lateral pinch the fingers flex a little bit more) it is plausible that the projections in the latent space are similar.

As already done with the SensorHand, we increased the complexity of the hand by assigning the hand 5 DoF – the flexion of each digit is actuated independently. As the thumb has an additional DoF (abduction/adduction) this value had to be specified as well. It was set into the intermediate rest position. Changing the value of this joint does not affect significantly the results.

The range of motion from each digit is sampled with 9 angles, resulting in the same number of points as in the SensorHand case. The resulting AI is 7.9%, which is considerably more than the 2.8% the original hand has. In that case, introducing additional DoFs is a suitable way to equip the hand with more anthropomorphic capabilities. If we observe the projection of the 5 DoF Michelangelo in Figure 12, we see that the extreme position in the right-left direction corresponds to the open and the grasp position respectively. All movements of the original Michelangelo lie beneath the line connecting those two positions and have very roughly a triangle shape. The movements of the 5 DoF hand overlap an additional space above that triangle. The top point in the projection in Figure 13 corresponds to a hand position where the index finger is extended but the other fingers are flexed and the thumb is in moderate flexion. The ability to individually flex fingers is important to reach new areas in the latent space. That

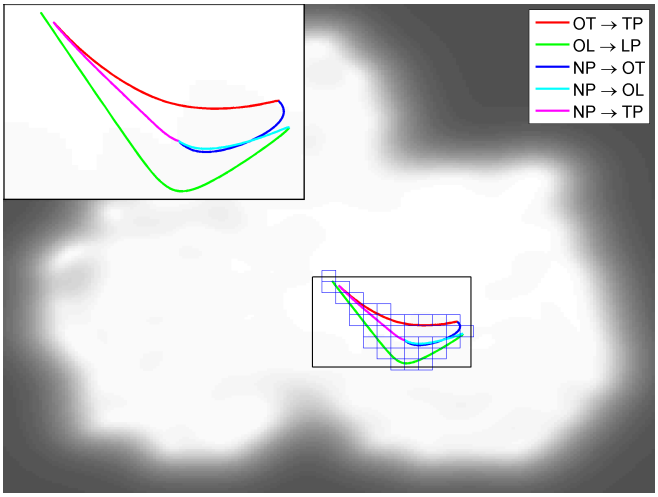


Fig. 12. Projection of the fingertip movements of the Otto Bock Michelangelo Hand to the latent space. The AI of the hand is 2.8%.

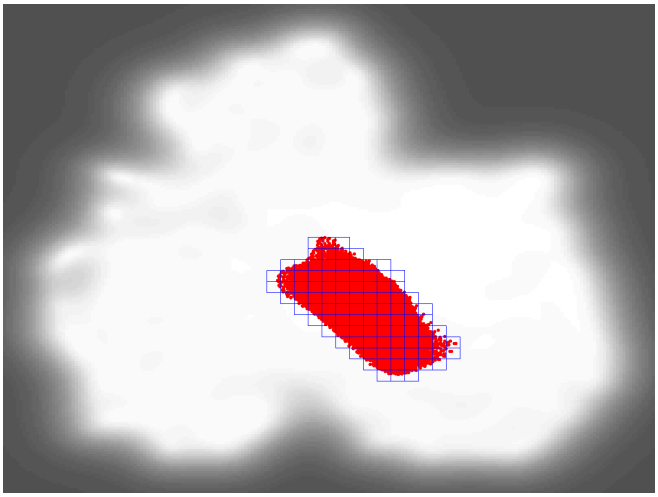


Fig. 13. Projection of the virtual 5 DoF Michelangelo hand to the latent space. The red points represent the area the hand can reach and it results in an AI of 7.9%.

is a difference with the SensorHand where the introduction of finger individuation does not influence substantially the latent space overlap.

C. FRH-4 Hand

As an example of a hand with many independent degrees of freedom, we use the FRH-4 hand [36] built for a mobile assisting robot ARMAR. With 8 independent fluidic actuators, it has a much more complex actuation system than the two prosthetic hands described in the previous sections. Its general appearance (Figure 14) is quite human-like; it has a size that is comparable to the human hand and the kinematic setup has some similarities. One design goal of the hand was to be anthropomorphic, but another goal was to develop a hand which is suitable for robotic grasping. To meet the second design objective, trade-off on the anthropomorphism had to be accepted. One major difference is the palm setup – the FRH-4 hand has one DoF in the metacarpus, which allows the palm

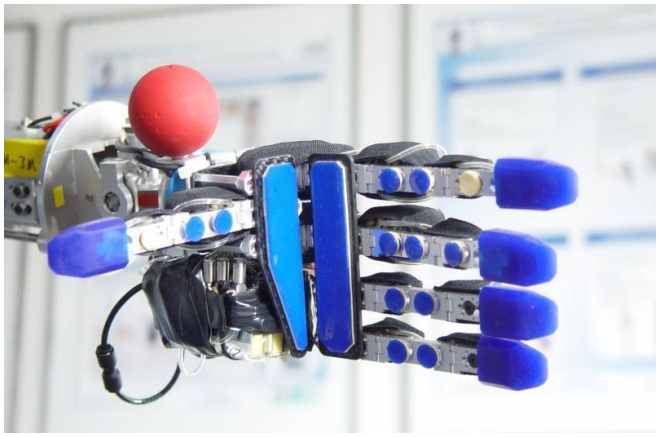


Fig. 14. FRH-4 hand [36].

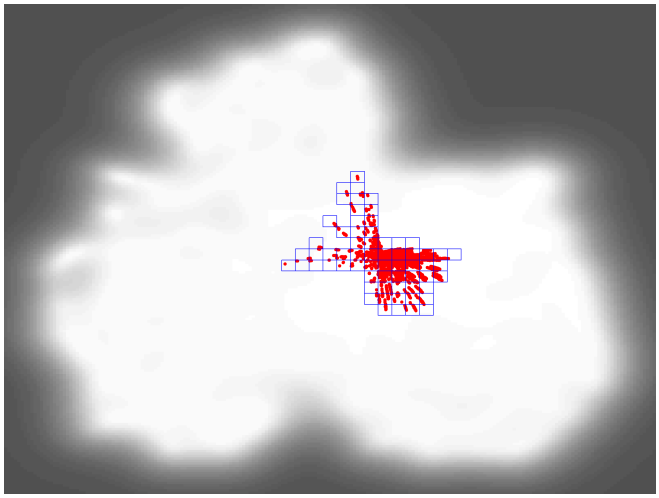


Fig. 15. Projection of the FRH-4 hand. It has an AI of 5.2%.

to flex in the middle. The human hand does not share this as the palm is rigid in the longitudinal direction. Figure 14 shows the palm joint in a flexed position, whereas in the extended position the fingers would point leftwards. The index and the middle finger both have 2 DoF, one joint represents the MCP joint of the human and the other one is in-between the PIP (Proximal interphalangeal) and DIP (Distal interphalangeal) joints. The ring and little fingers have one combined DoF, that is a common flexion in the MCP joint. All joint axes of the fingers are parallel and the finger segment lengths are 40 mm. The thumb has two actuators, which actuate the CMC joint and the joint between the MCP and IP (Interphalangeal) joint of the thumb. The base of the thumb is exactly opposing the index and the middle fingers. This setup is very similar to the SensorHand and substantially different to the human hand where the axes of the thumb are not aligned with the axes of the fingers.

Each of the eight DoF has a range of 90 degrees and to calculate all hand configurations we took four samples from each of the joint workspaces. Each joint can be flexed by $\{0, 30, 60, 90\}$ degrees and due to the high dimensionality of the hand this leads to a total number of $4^8 = 65536$ hand configurations. Further increasing the number of samples would require prohibitively large computational times.

As the kinematic structure makes it difficult to define where the hand length could be measured (which is used for scaling the positions prior to projecting to the latent space), we performed a parameter sweep through all possible hand lengths and then determined the hand length with the maximal overlap. This length was assumed to be the correct hand length and the results corresponding to that length are given. For the two prosthetic hands we do not have to calculate the hand length, as there is information on the size available. The resulting hand length of the FRH-4 hand is 25 cm, which is slightly larger than the maximal human hand length of about 21.15 cm. [37]. The width of the FRH-4 hand is 9.3 cm [36] which is comparable to the hand width of a large human hand [37]. The calculated hand length is slightly too large given the hand width but due to the different kinematic setup to the human hand, that difference is acceptable.

The projection of the hand with the determined hand length of 25 cm is shown in Figure 15. The anthropomorphism index is 5.2%. Compared to the large number of actuators this is a relatively low value, given that the Michelangelo hand with only 2 DoF has already an AI of 2.8%. As described above, the hand has some features which are not anthropomorphic, which explains the reduced score.

In Figure 15 we observe that the outermost points are slightly isolated from the rest. That violates the guideline on the joint space sampling, as the inter-point distances in the latent space should be smaller than the box size. Consequently only a few points are located at the intersection with the human action manifold. If we increase the number of hand configurations, we would have been able to further increase the overlap as those points would not be isolated anymore. To test how much larger the overlap could be, we exchanged the way to obtain the joint values. Instead of a systematic variation of the joint angles, we sampled the joint space with 60000 random points and calculated the corresponding overlap for five such sets. The result was an AI of $9.2 \pm 0.25\%$. The different sampling method increased the overlap, but is still small when compared to the human hand. For the SensorHand and the Michelangelo Hand this resampling was not necessary as their joint space could be sampled densely enough.

VIII. Discussion and Conclusions

We have presented a methodology for measuring the differences between human and artificial hand capabilities. The similarity of an artificial hand with respect to the human hand is determined by the definition of an anthropomorphism index (AI). We concentrate specifically on evaluating the capability of the hands to execute different grasping actions. Human hand data is generated from five test subjects and artificial hands data is generated by sampling their joint space and calculating the corresponding fingertip poses via forward kinematics. The contribution of the work is the first attempt to develop a metric for a comparison based on state of the art methods for nonlinear dimensionality reduction.

The big advantage of the system is that it offers great flexibility with respect to the hands that can be tested. The

hands can have an arbitrary kinematic structure and the joint couplings can be very complex. The method can easily be used for other similar purposes: it only requires to generate a new underlying dataset. For example, if one wants to emphasize precision grasps, we could record humans grasping a variety of small objects.

The AI evaluates an important proportion of the hand capabilities, which are its kinematics. There are also other parameters that are of relevance for a functional end-effector, those are for example speed, precision and force of the hand. Most of those parameters are connected to the mechanical implementation, whereas the AI evaluates the underlying kinematic setup.

There are lessons learned in terms of the employed methodology. As the inter-point distances on the human manifold have some certain average value, the kernel width has to compensate for this. If this is not taken into account, the manifold cannot be represented properly as there may be holes where the projection is not supported by data. This defines a minimal kernel length which can represent the manifold properly and also introduces a minimal selectivity perpendicular to the manifold. The sampling rate of the human dataset can be increased in order to improve the selectivity, but at the cost of higher computational requirements in terms of memory and processing capabilities.

As there are no objects involved in assessing the anthropomorphic hand structures, passive compliance and underactuated hands cannot be implemented directly. Interaction with the object is needed to determine how the fingers wrap around it. In order to analyze hands with passive compliance, a workaround can be used by sampling the passive joints as well. This might not deliver the most accurate results, but it will provide hints on the capabilities of such a hand.

The experimental evaluation shows that hands with as little as two actuators (like the Michelangelo hand) are able to populate large proportions of the latent space of the low-dimensional human hand movements. Various studies ([9], [38]) have shown that human hand movements can be described with high accuracy in a linear subspace of eight dimensions. According to those studies, having a hand with eight DoFs or less should be sufficient to cover most of the human hand movement.

In general, the tested hands covered a relatively low area with an overlap of less than 10%: thus, the hands had significant limitations compared to the human hand. Some of the hands are not able to fully extend and flex the fingers due to rigid fingers (SensorHand and Michelangelo), having joint axes that are not well aligned with the movement axis of the human hand (SensorHand and FRH-4 Hand) or having a range of motion in the joints that is lower than the humans (SensorHand and Michelangelo). Those are the reasons for the reduced overlap and the goal for the future is to overcome these limitations with as little effort (actuators, joints, etc.) as possible.

As the next step, we plan to perform parameter studies using the system to determine the influence of design parameters on the AI. That should provide insights about the relationship between kinematic elements and their influence on grasping capabilities. The final goal is not only to change parameters, but to find the optimal kinematic structure with respect to the proposed anthropomorphism index.

REFERENCES

[1] M. T. Mason, S. Srinivasa, A. S. Vazquez, and A. Rodriguez, "Generality and simple hands," Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-10-40, November 2010.

[2] L. Biagiotti, F. Lotti, C. Melchiorri, and G. Vassura, "How far is the human hand? a review on anthropomorphic end effectors," DIES Internal Report, University of Bologna, Tech. Rep., 2004.

[3] M. C. Carrozza, G. Cappiello, S. Micera, B. B. Edin, L. Beccai, and C. Cipriani, "Design of a cybernetic hand for perception and action," *Biological Cybernetics*, vol. 95, no. 6, pp. 629–644, Dec. 2006.

[4] C. S. Lovchik and M. A. Diftler, "The robonaut hand: a dexterous robot hand for space," in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, vol. 2, 1999, pp. 907–912 vol.2.

[5] H. Liu, K. Wu, P. Meusel, N. Seitz, G. Hirzinger, M. H. Jin, Y. W. Liu, S. W. Fan, T. Lan, and Z. P. Chen, "Multisensory five-finger dexterous hand: The DLR/HIT hand II," in *IROS*, Sep. 2008, pp. 3692–3697.

[6] H. Rijpkema and M. Girard, "Computer animation of knowledge-based human grasping," *SIGGRAPH Comput. Graph.*, vol. 25, no. 4, pp. 339–348, 1991.

[7] B. Buchholz and T. J. Armstrong, "A kinematic model of the human hand to evaluate its prehensile capabilities," *Journal of Biomechanics*, vol. 25, no. 2, pp. 149 – 162, 1992.

[8] D. Dragulescu, V. Perdereau, M. Drouin, L. Ungureanu, and K. Menyhardt, "3d active workspace of human hand anatomical model," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 15, 2007.

[9] J. S. M. Santello, M. Flanders, "Postural hand synergies for tool use," in *The Journal of Neuroscience*, 1998.

[10] C. R. Mason, J. E. Gomez, and T. J. Ebner, "Hand synergies during reach-to-grasp," *J Neurophysiol*, vol. 86, no. 6, pp. 2896–2910, December 2001.

[11] I. V. Grinyagin, E. V. Biryukova, and M. A. Maier, "Kinematic and dynamic synergies of human precision-grip movements," *Journal of Neurophysiology*, vol. 94, no. 4, pp. 2284–2294, 2005.

[12] J. Romero, T. Feix, H. Kjellström, and D. Kragic, "Spatio-temporal modeling of grasping actions," in *IROS*. IEEE, 2010.

[13] F. Lotti, P. Tiezzi, G. Vassura, L. Biagiotti, G. Palli, and C. Melchiorri, "Development of UB hand 3: Early results," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, 2005, pp. 4488–4493.

[14] A. Bicci, "Revisiting grasping basics," in *Robotics, Science and Systems Conference: Workshop Grasp Acquisition: How to Realize Good Grasps*, Jun. 2010.

[15] A. Bicchi, "Hands for dexterous manipulation and robust grasping: a difficult road toward simplicity," *Robotics and Automation, IEEE Transactions on*, vol. 16, no. 6, pp. 652–662, 2000.

[16] G. A. Kragten and J. L. Herder, "The ability of underactuated hands to grasp and hold objects," *Mechanism and Machine Theory*, vol. 45, no. 3, pp. 408–425, Mar. 2010.

[17] M. Ciocarlie and P. Allen, "A design and analysis tool for underactuated compliant hands," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*. IEEE, October 2009, pp. 5234–5239.

[18] ——, "Data-driven optimization for underactuated robotic hands," in *2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*. IEEE, May 2010, pp. 1292–1299.

[19] D. Pratichizzo, M. Malvezzi, and A. Bicchi, "On motion and force controllability of grasping hands with postural synergies," in *Proceedings of Robotics: Science and Systems*, Zaragoza, Spain, June 2010.

[20] G. A. Kragten and J. L. Herder, "A platform for grasp performance assessment in compliant or underactuated hands," *Journal of Mechanical Design*, vol. 132, no. 2, 2010.

[21] M. Malhotra and Y. Nakamura, "The relationship between actuator reduction and controllability for a robotic hand," in *IEEE International conference on Biomedical Robotics and Biomechatronics*, 2010.

[22] W. Craelius, "The bionic man: Restoring mobility," *Science*, vol. 295, no. 5557, pp. 1018–1021, Feb. 2002.

[23] P. Parker, K. Englehart, and B. Hudgins, "Myoelectric signal processing for control of powered limb prostheses." *Journal of electromyography and kinesiology*, vol. 16, no. 6, pp. 541–548, Dec. 2006.

[24] P. J. Kyberd, C. Wartenberg, L. Sandsjö, S. Jönsson, D. Gow, J. Frid, C. Almström, and L. Sperling, "Survey of upper extremity prosthesis users in sweden and the united kingdom," *JPO: Journal of Prosthetics and Orthotics*, vol. 19, no. 2, p. 55, 2007.

[25] R. N. Rohling and J. M. Hollerbach, "Optimized fingertip mapping for teleoperation of dextrous robot hands," pp. 769–775.

[26] M. Fischer, P. van der Smagt, and G. Hirzinger, "Learning techniques in a dataglove based telemanipulation system for the DLR hand," pp. 1603–1608.

[27] H. Hu, X. Gao, J. Li, J. Wang, and H. Liu, "Calibrating human hand for teleoperating the hit/dlr hand," in *IEEE International Conference on Robotics and Automation*, 2004, pp. 4571–4576.

[28] N. D. Lawrence and J. Quinonero-Candela, "Local distance preservation in the gp-lvm through back constraints," in *ICML06*, 2006, pp. 513–520.

[29] R. M. Murray, S. S. Sastry, and L. Zexiang, *A Mathematical Introduction to Robotic Manipulation*. Boca Raton, FL, USA: CRC Press, Inc., 1994.

[30] P. E. Nikravesh, *Computer-Aided Analysis of Mechanical Systems*. Prentice Hall, 1988, vol. 186.

[31] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, June 2009.

[32] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[33] P. Corke, "A robotics toolbox for MATLAB," *IEEE Robotics and Automation Magazine*, vol. 3, no. 1, pp. 24–32, Mar. 1996.

[34] Otto Bock. (2011) SensorHand speed. [Online]. Available: http://www.ottobock.com/cps/rde/xchg/ob_com_en/hs.xsl/3652.html

[35] ——. (2011) Otto Bock at the trade fair 2010 Leipzig. [Online]. Available: http://leipzig.ottobock.de/index.php?id=161&no_cache=1&L=1

[36] I. Gaiser, S. Schulz, A. Kargov, H. Klosek, A. Bierbaum, C. Pylatiuk, R. Oberle, T. Werner, T. Asfour, G. Bretthauer, and R. Dillmann, "A new anthropomorphic robotic hand," in *2008 8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008)*. IEEE, Dec. 2008, pp. 418–422.

[37] J. W. Garrett, "Anthropometry of the hands of male air force flight personnel," Aerospace Medical Research Laboratory, Aerospace Medical Division, Air Force Systems Command,, Wright-Patterson Air Force Base, OH, USA, Tech. Rep., 1970.

[38] J. Ingram, K. Körding, I. Howard, and D. Wolpert, "The statistics of natural hand movements," *Experimental Brain Research*, vol. 188, no. 2, pp. 223–236, June 2008.

Extracting Postural Synergies for Grasping

Javier Romero, Thomas Feix, Carl Henrik Ek, Hedvig Kjellström and Danica Kragic

***Abstract*—Observation and analysis of human motion is often used for planning and control of human inspired movements in robots. This includes examples of arm/hand movements and gait control. Human data is usually high-dimensional and in many cases it is used to control a robot which much fewer degrees of freedom. To that end, different representations based on dimensionality reduction techniques have been used to enable viable control solutions. In control and planning of grasping movements in particular, postural synergies have been used as a low-dimensional representation to enable establishing correspondence between human and robot hand activities. In their original formulation, postural synergies are based on linear dimensionality reduction methods that, as we will show in this paper, do not represent human hand activity with sufficient accuracy due to inherit non-linearities in the data. Thus, the work presented in this paper addresses non-linear dimensionality reduction methods and their application to human hand data.**

In addition to adressing encoding of postural synergies, our work relates closely to recent work in robotic control of combined reaching and grasping movements. However, this work is based on an assumption that correlations in the data is evidence of causal relation, an assumption that may not hold. Non-linear dimensionality reduction methods may be used to tackle the correlations problem not by considering causal relations between dimensions, but by considering them being generated from an external manifold which has to be inferred. Showing how this can be done is the first contribution of our work. Another strong contribution of this paper is the analysis of the internal parameters used in dimensionality reduction techniques, which sheds light into algorithms which have been traditionally used as a "black-box" in robotics. Finally, we provide a thorough experimental evaluation that shows how the proposed methods outperform the standard techniques in the field both in terms of recognition and generation of motion patterns.

I. Introduction

Control of reaching and grasping movements in robots relies often on the analysis of human data, [1], [2], [3], [4], [5], and considers problems from data representation to planning and mapping, see Fig. 1. The analysis of human grasping has been widely studied in neurophysiology and psychology, where the goal is to understand the processes behind the control of movement, [6], [7], [8], [9]. A central result of such studies is the evidence of very strong correlations in the finger positions and their movements, implying a large redundancy. The existence of those correlations allows us to extract compact representations, often referred as *synergies* (e.g. trajectories in a lower dimensional space $\mathbf{X}$ in Fig. 1), which can concisely describe complex, but redundant, human motion e.g. trajectories in high-dimensional space $\mathbf{Y^h}$ in Fig. 1. Grinyagin et al. [9] classifies synergies into three types. First, static postural synergies, that refer to the correlation between single kinematic poses, e.g. [7]. Second, kinematic synergies that consider time dependent correlation of postures during an action, e.g. [8]. Finally, muscle synergies address the covariation of lower level representations of movement such as electromyographic activity, [10], [6]. While muscle synergies are specifically bound to human because of their internal nature, postural synergies and kinematic synergies have inspired a large body of work in robotics, [7], [11].

In regard to modeling of the synergies, most of the work in neurophysiology uses linear models for encoding postures and movement: the three commonly cited studies [7], [8], [9] use linear techniques. Even comparative studies of techniques used for synergy extraction like [12] employ linear methods only. In this paper we show that, apart of being high-dimensional, finger movements are non-linear. Therefore, a more natural approach is to encode them using techniques that take not only high-dimensionality into account, but also non-linearity.

Recent work of [3] states that the reasons why reaching and grasping movements have not been addressed in a proper manner in robot control is exactly their complexity (related to their non-linear nature) and high dimensionality.

In neurophysiology, it is argued that low dimensional representations (i.e. synergies) drive the modulation of muscle forces for control of human posture, [6]. Therefore, synergies can be used as a modeling paradigm in robot control, where control laws in low dimensional space $\mathbf{X}$ can drive the forces applied on the higher dimensional robot space $\mathbf{Y}^r$, Fig. 1. For example, they have been used to design reference robot hand movements that adapt to external forces (originating from objects) on demand, [11]. The use of synergies makes the reference movements lower dimensional, with the obvious advantages that the lower dimensionality conveys.

Another closely related area that suffers from the curse of dimensionality is planning. The computational cost of searching for adequate kinematic configurations increases exponentially with the dimensionality of such configurations. In [13], the authors reduced such complexity by searching for good grasping postures in a postural synergy space ($\mathbf{X}$ in Fig. 1) of lower dimensionality: grasps are planned for various robotic hands based on the grasping synergies extracted from human data.

Grasping synergies have also been used as a common, semantic representation that transcends differences in embodiments and attacks the so called "correspondence problem", [14]. This problem, also known as *Mapping*, refers to transferring postures or movements from one agent to another (e.g. mapping human postural space $\mathbf{Y}^h$ to robot space $\mathbf{Y}^r$, Fig. 1). While the correspondence is designed manually in synergy space in [13] (correspondences eigengrasp-robot $\mathbf{X} \rightarrow \mathbf{Y}^r$ are assigned manually), Kang et. al. apply similar concepts (a simpler, semantic representation of grasping poses known as *Virtual Fingers*) to compute such a mapping automatically, [15].

In summary, the efficient representation provided by postural synergies have been used to address the inherent problems

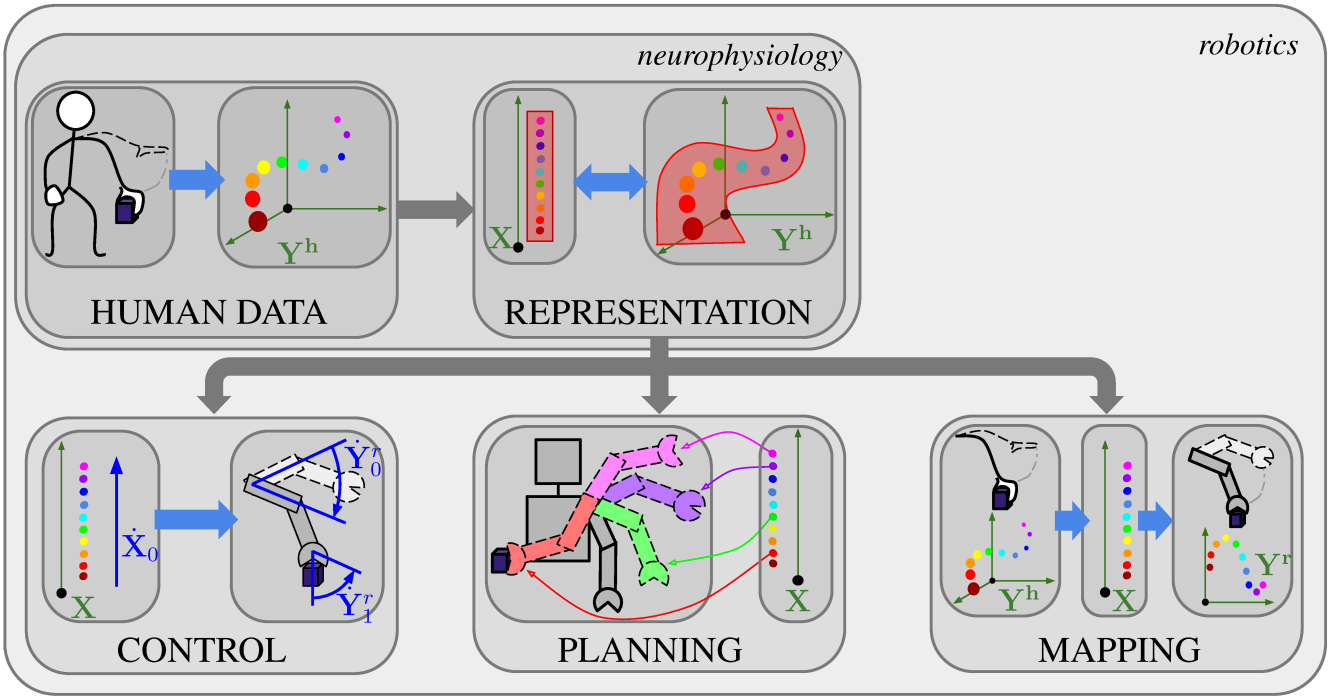


Fig. 1: The study of human data and its representation has been used in neurophysiology for studying how the brain and the nervous system command motor actions. Robotics has benefited from such studies and expanded them. The results of those studies have been used for control of dynamic movement in robots (Control), search for suitable poses for a specific task (Planning) and transferring kinematic actions between different embodiments (Mapping).

related to the high-dimensionality of the above applications. The central question arises: *To what extent can we rely on the linear encoding of human data?*

To better motivate the problem, we show how a recent, state-of-the-art work in robotics, such as [3], can benefit from using a more appropriate representation of synergies. Shukla and Billard [3] improve the robustness to perturbations of grasping actions by exploiting correlations between the reaching and grasping components. The system shows how the coupling between reaching and grasping acts as an "attractor" which allows the task to be perturbed in terms of goal position or grasp type. This is conceptually similar to the role of synergies in [11], which can be perturbed by external forces.

More specifically, [3] addresses the correlations and coupling in reaching and grasping movements. The approach is based on the assumption that there is a causal relationship between two processes (reach and grasp). This implies that the system is *a-priori* divided into two parts and the role of master and slave are manually assigned to them. While this may seem reasonable in the case of reaching and grasping, in other cases (e.g. correlations between fingers in a manipulation task) it is not so easy to reason about what and where the correlation exists: i.e. Does moving the thumb cause the movement of the pinky or vice-versa?

More important, causal relationships easily break down when the task becomes more complex. For example, extending the reaching/grasping with a retreat movement breaks the causal relationship between reaching and grasping, because closeness to the body does not *cause* large hand aperture any more the relation between reach and grasp is no longer functional (i.e. $\mathbf{y} = f(\mathbf{x})$ is unique,[16]), and therefore cannot be causal (see Fig. 2,middle). Another problem of the causal relation between master and slave is that it only relates two variables or sets of variables. Imagine that the dataset to model is a path embedded on a sphere sector in three-dimensional space $\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{Y}_2$, rightmost Fig. 2. According to [3], multiple variables can be grouped into master or slave. This approach implies that one of the dimensions generate the other two, or vice-versa. However, that model misses important correlations within the two-dimensional set.

Dimensionality-reduction methods tackle the correlations problem not by considering causal relationships between dimensions, but by considering them being generated from an external manifold which has to be inferred ($\mathbf{X}$ in Fig. 2). Such manifolds define the synergy space. As mentioned previously, synergies have traditionally been extracted with linear methods which also impose a functional form (linear in this case) of the mapping. Our first contribution is the employment of recently developed non-linear dimensionality techniques in the context of robotics. Such methods avoid the manual design of causal relations in [3], and overcome the limitations of traditional techniques inherent to their linear nature. Another strong contribution of this paper is the analysis of the internal parameters used in dimensionality-reduction techniques, which sheds light into algorithms which have been traditionally used as "black-boxes" in robotics. Finally, we provide a thorough experimental evaluation that shows how the proposed methods outperform the standard techniques in the field both in terms

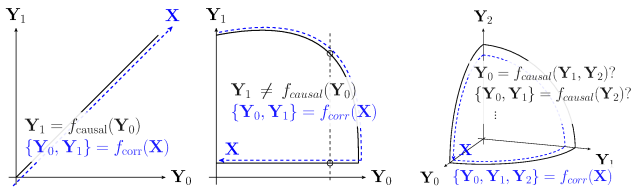


Fig. 2: Applicability of causal relation between data dimensions. $\mathbf{Y}_i$ represent robot dimensions such as finger joints, or gripper aperture vs distance to the object, while $\mathbf{X}$ represents a (non-linear) synergy that can model the correlations on $\mathbf{Y}_i$. The master-slave relation between different $\mathbf{Y}_i$ is only applicable in simple cases (left figure); such relation is not possible when the relation between variables is not functional (middle figure), because points with the same coordinate $\mathbf{Y}_0$ (e.g. arm extension) have different coordinate $\mathbf{Y}_1$ (e.g. hand aperture). Moreover, in higher dimensional cases the identification of master-slave relation becomes harder and more inefficient.

of recognition and generation of motion patterns.

The paper is organized as follows. In Section II, we reviewed the existing work in terms of synergy extraction in robotics and provide motivation why current linear approaches are not sufficient. The remaining sections describe the methodology proposed in this paper. Section III describes different dimensionality-reduction methods. The human data used in this paper is described in Section IV. We qualitatively analyze the extracted grasping models in Section V, and evaluate their performance in Section VI. Finally, we conclude the article in VII.

II. Related Work and Motivation

There has been a substantial amount of research directed towards using synergies in human and robot grasping. Commonly, synergetic representations are extracted using linear dimensionality-reduction methods such as Principal Component Analysis (PCA) [17]. One of the earliest appraoches is presented in Santello et al., [7] The authors showed that a substantial part of the variations of grasping hand poses (80% of the data *variance*) can be expressed as a two-dimensional linear combination of hand joints.

While Santello el al. showed the correlation of joints for different static hand poses, later research focused on the temporal correlation (i.e. kinematic synergies) of the hand pose while executing specific grasps. In [9], multiple executions of precision grips are analyzed with PCA to conclude that a one-dimensional direction can explain more than 97% of the data variance. The data from six different subjects and three different grasping conditions was analyzed separately, generating a different one-dimensional manifold for each of these series of twenty trials. Mason et al. [8] studied the correlation in the position of different parts of the hand for specific grasps applied to different objects, using again similar techniques as Santello et al. They concluded that for each subject and grasp, more than 96% of the variance could be explained by a one-dimensional manifold.

The concept of synergies have had a large impact in research on robotics. Ciocarlie et al. [13] introduced the concept of Eigengrasps based on the grasp synergies defined in [7]. In this system, grasps are planned in terms of Eigengrasps instead of manipulator degrees of freedom, making the optimization of the hand pose more computationally tractable. Eigengrasps were also used in [4] to control the 12 degrees-of-freedom of a robotic hand. In [18] the grasping control of a 17-dimensional hand was performed by moving on a 2-dimensional manifold extracted with Isomap. In [19], correlations between wrist and finger movements were modeled, validated and applied to solve control of redundant degrees of freedom. There is also some work focused on imposing certain desired characteristics on the synergies. Steffen et al. [20] impose that the low-dimensional space variables represent task evolution (time) and relevant parameters for the action (such as object size). Bitzer et al. [21] interpolates between high-dimensional actions by imposing certain spatial relations between the training actions in a low-dimensional representation which resembles the task-space. Another area of robotics influenced by the concept of synergies is robotic hand design, [22], [23], [2].

The majority of the mentioned systems rely on linear dimensionality-reduction methods. Although it is clear that the linear approximation of the mapping can result in inaccuracies, it is commonly believed that a linear subspace recovered by PCA represents the motion correlations with sufficient accuracy. However, we argue that the relation between kinematic dimensions is highly non-linear, and those non-linearities are critical for a proper representation of the correlations. This is because linear analysis of such correlations can be misleading. A small part of the variance can hold critical details of a particular action as we will demonstrate below.

In summary, linear dimensionality reduction such as PCA is fast, robust and conceptually simple. However, its results deteriorate rapidly when the data becomes non-linear. This is problematic when non-linearities in the data are crucial for the success of the task. For motivational purposes, let us depict a concise example of how linear models that properly represent most of the data variance can fail to represent the motion correlation in an action such as reaching and grasping.

Our example consists of a set of robot grasping executions for which we wish to acquire a low dimensional representation. For illustrative purposes, our robot will be fairly simple. It has two components: a telescopic arm and a gripper, each with one degree of freedom (see Fig. 3). Our goal is to obtain a one-dimensional representation of a grasping task. Despite its simplicity, this example shows how non-linearities in high dimensional space can be critical in robot control, making linear synergies unsuitable for robot-control tasks.

Grasping an object can be roughly divided into two components: transport phase and grasping phase. During the transport phase, the hand slowly reaches its maximum aperture (considering the hand being initially closed) and in the grasping phase the hand closes very fast, while the distance to the object decreases only slightly. The following equation formalizes this idea:

$$y^r_{t,1} \propto (1 - y^r_{t,0})^{\frac{1}{3}+\epsilon_0} + \epsilon_1 \quad (1)$$

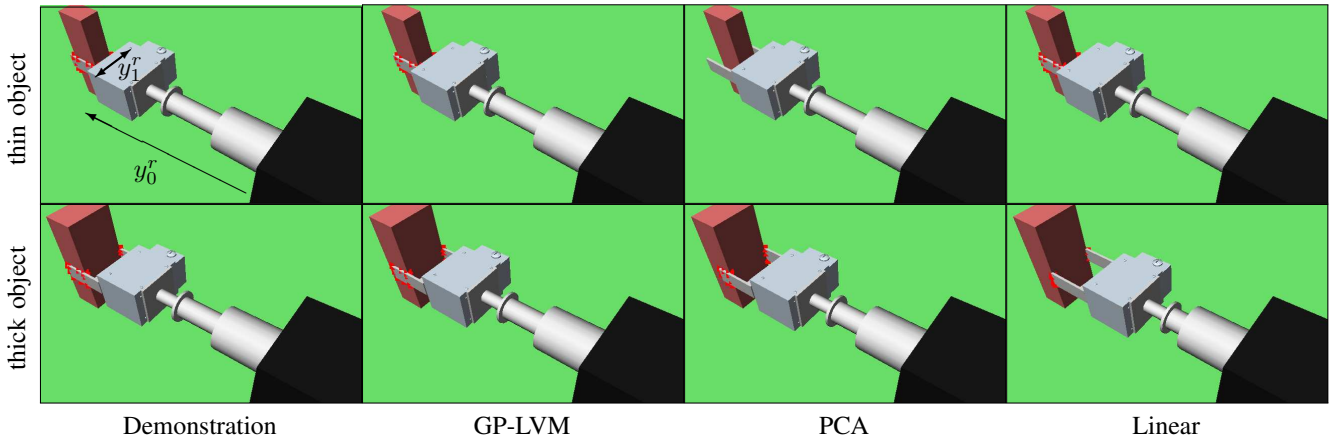


Fig. 3: Simple grasp execution. The columns represent, from left to right, one of the original examples, GP-LVM, PCA and Linear models. The GP-LVM models mimics almost perfectly the training example. PCA fails to make contact for the thin object, and grasps the thick object slightly prematurely. The Linear model grasps properly the thin object, but collides with the thick object. Graphs generated with OpenRave, [24]

where $\mathbf{y_t^r}$ represent the two dimensional configuration of the robot at every time, $y_{t,1}^r$ represents the gripper aperture at time t, $y_{t,0}^r$ the arm extension at time t, and ϵ_i represent different noise terms. Ten grasping sequences $\{y^r\}$ were generated, and three models were extracted from them. The first one, tagged as "Linear" in top Fig. 4, is a linear interpolation between the maximum and minimum aperture with linear decrease of arm length. The other two are the trajectories generated by sampling a one-dimensional space extracted from the data using PCA and GP-LVM.

As it could be expected, the non-linear method, GP-LVM, resembles best the original trajectories which were highly non-linear. Let us consider the consequences of the linearization. In Fig. 4, when the linear trajectories are below the example trajectories (all the Linear model, and part of the PCA one) the gripper is more closed than in the original examples. That means that the gripper can potentially collide with the object (see Linear column, thick object row in Fig. 3). When the line is above the examples (last stage of PCA model) the gripper is more open than in the examples, and therefore the gripper might fail to enter contact with the object (PCA column, thin object row in Fig. 3).

The relation between the two joints in this action set was bijective, and the PCA reconstruction is relatively good, even though it fails in some cases. If the actions also includes the retreat movement (retract the arm with the gripper set to a constant aperture) the relation becomes more non-linear (more difficult to be modeled by PCA) and non-bijective (not possible to be modeled as a causal relation between master and slave, [3]), bottom Fig. 4. The linear manifold computed by PCA averages the approach and retreat actions, setting a gripper aperture which is too small for approaching the object. But most importantly, it cannot retreat with a constant aperture.

The relation between the transport and grasp components in a grasping action is inherently non-linear. Trying to model it as a linear manifold can result in early collisions or fail to contact the object. This can happen even when a large part of the variance is kept in the linear manifold. PCA managed to represent 98.6% of the variance of the grasp-approach action with its linear representation, and still fails to grasp properly the object. The amount of variance represented by PCA for the grasp-approach-retreat action was also high, 78.6%.

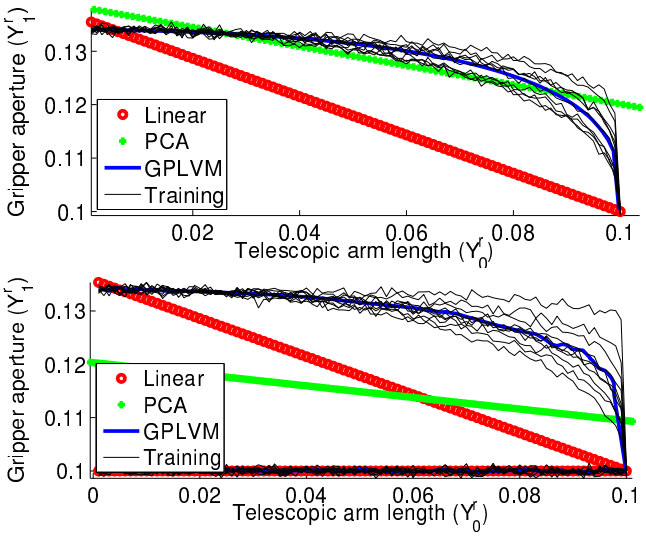


Fig. 4: Grasp trajectories generated from different models computed from the ten executions shown as *Training*. *Linear* is a linear interpolation between the beginning and end gripper aperture. *PCA* and *GP-LVM* extract a one-dimensional representation of data, and generate the trajectory by linearly sampling that one-dimensional manifold. Top figure shows only approach and grasp, bottom figure includes retreat.

The following sections of this paper will further explain the principles that drive PCA and GP-LVM, focusing on their assumptions about the data and the parameters that can tune their performance.

III. Methods for Extracting Postural Synergies

In this section we will motivate and formulate the problem of dimensionality-reduction. We will then proceed to introduce the two diametrically different approaches we will evaluate in this paper.

For many scenarios, data is observed in a representation that is significantly different from the intrinsic representation of the data. In specific, this often implies that the observed representation is an "over" representation of the data in terms of degrees-of-freedom because the data actually lies on or close to a lower dimensional manifold in the observed representation. The task of dimensionality reduction is to, given data in a specific representation, recover the intrinsic representation. The problem is formalised as follows. Given a set of data $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ where $\mathbf{y}_i \in \Re^D$ we assume this to have been generated from a intrinsic representation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, $\mathbf{x}_i \in \Re^q$ through the generative mapping f,

$$\mathbf{y}_i = f(\mathbf{x}_i). \tag{2}$$

Further, we will assume the observed representation to be an over parametrization implying that $q < D$. The objective of dimensionality reduction is to recover $\mathbf{X}$ from $\mathbf{Y}$.

The problem is severely ill-constrained since an infinite combination of input representations $\mathbf{X}$ and mappings f could have generated $\mathbf{Y}$. Different algorithms make different assumptions in order to proceed. There are two main branches of work in dimensionality-reduction, *spectral* and *generative*. Spectral approaches assume the generative mapping f to have a smooth inverse according to some metric. This is different compared to the generative class of models which directly tries to model the generative mapping. The spectral assumption is stronger and does therefore constrain the solution space further compared to the generative. This implies that while the generative models are applicable to a larger range of data, recovering the solution might be a significant challenge. There are both linear and non-linear formulations of the methods.

In this paper, we mainly focus at two different algorithms, Principal Component Analysis (PCA) and Gaussian Process Latent Variable Models (GP-LVM). PCA is a spectral linear model while GP-LVM is generative and capable of modelling a non-linear generative mapping. Our motivation of evaluating the performance of these two methods stems from the fact that PCA has been the dominant algorithm for extracting postural synergies while the GP-LVM is one of the most recently proposed and flexible algorithms in the area of machine learning. We also consider the usage of two spectral non-linear methods such as Isomap and Locally Linear Embeddings (LLE). We do not evaluate these methods in depth because, as we will show, initial results showed they are not suitable for our task. The reminder of this section will describe the different methods and further motivate our choice.

A. Principle Component Analysis

Principle Component Analysis (PCA) is a method for dimensionality reduction frequently applied to a large range of applications. The objective of the algorithm is to find a low-dimensional hyper-plane which maximizes the variance of the data projected onto it (i.e., which minimizes the reconstruction error of the data). Mathematically, this implies finding a low-rank approximation of the covariance matrix of the data which minimizes the Frobenius norm,

$$\mathrm{E}(\mathbf{C}) = ||\mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{C}||, \tag{3}$$

where $\mathbf{Y}$ is the centered observed representation of the data and $\mathbf{C}$ the approximation to the covariance matrix. The optimal solution to Eq. 3 can be found in close form through the eigendecomposition of the covariance matrix in the observed space,

$$\mathbf{C} = \sum_{i=1}^{D} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}, \tag{4}$$

where λ_i and $\mathbf{v}_i$ are the $i : th$ eigen value and vector of the eigendecomposition of the covariance matrix in the observed space. $\mathbf{v}_i$ specifies an orthonormal basis. Therefore, the "mass" provided by each component is proportional to the corresponding eigenvalue. The best rank k approximation of the covariance matrix is then computed based on the eigenvectors correspondent to the largest k eigenvalues,

$$\hat{\mathbf{C}}_k = \mathrm{argmin}_{\mathbf{C}}\ \mathrm{E}(\mathbf{C}) = \sum_{i=1}^{k} \lambda_i \mathbf{v_i} \mathbf{v_i}^{\mathrm{T}} \qquad \lambda_i \geq \lambda_j, \quad i < j. \tag{5}$$

Consequently, the reconstruction of the data takes the form of the linear projection $f(\mathbf{X}) = \mathbf{V}_{1\to k}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{V}_{1\to k}^{\mathrm{T}}$ represents the transpose of a matrix composed by the k eigenvectors with highest eigenvalues. The major benefit of the algorithm is that it is robust compared to methods such as Isomap or LLE, as it relies on global statistics of the data. If it can be assumed that the noise in the data is of low variance and that the intrinsic signal occupies a linear subspace in the observed representation, then it will recover the correct space. However, these are strong assumptions that do not apply to many types of data as a significant portion of the observations corresponds to noise and/or the correlations in the data are non-linear.

Another way of interpreting PCA optimization is that it minimizes the variance of the data projected to directions perpendicular to the extracted hyperplane $\mathbf{V}_{1\to k}$. However, variance of different data dimensions obviously depends on the general scale of this dimensions; for example, in a grasping action, the position of the wrist relative to the chest varies more than the position of a finger relative to the wrist. This means that the error of high varying dimensions will shadow the one from low varying dimensions. We can control the contribution of each dimension to the error function by pre-scaling the data. If prior knowledge is available, this can be exploited by transforming the data such that the L_2 distance better reflects our notion of similarity. Since we do not want to commit to any special purpose scaling for the grasping data we have

collected, we scale each dimension to make it have the same variance; this process is called "whitening".

B. GP-LVM

Spectral approaches such as PCA aims to find a transformation that maps from the observed to the intrinsic representation. Underlying such an approach is an assumption that the generative mapping is invertible (e.g. linear mappings in PCA). Generative models do not rely on this assumption and directly models the generative mapping. However, this also means that the solution space is much less constrained. In order to proceed we need to somehow be able to rank or rate different possible solutions. The probabilistic approach is to formulate the likelihood, of the underlying model to have generated the observed data. Assuming the observed data has been generated from the intrinsic through a mapping $f(\cdot)$ corrupted by Gaussian noise $\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I})$ the likelihood of the data can be formulated as follows,

$$p(\mathbf{Y}|\mathbf{f}, \sigma^{-2}) = \prod_i \mathcal{N}(\mathbf{y}_i|\mathbf{f}_i, \sigma^{-2}), \tag{6}$$

where σ is the noise variance and $\mathbf{f}_i$ the instanciations of the mapping. In the generative framework, the intrinsic representation $\mathbf{X}$ will be referred to as the *latent* representation of the data. The likelihood function allows us to evaluate how likely it is that a specific model $f(\cdot)$ has generated the observed data. However, the solution space is still enormous (and subject to local minima) and in order to find a reliable solution regularization is needed.

The Bayesian approach to proceed is to formulate a prior that encodes a preference toward certain solutions and combines this with the likelihood to regularize the problem. In probabilistic PCA [25] the author places a prior over the latent locations $\mathbf{X}$ while the maximum likelihood solution to the parameters of the mapping f are found. A different approach is to place the prior over the mapping (i.e. considering all the possible mappings and integrating over them). This can be done in a flexible non-parametric manner by using a Gaussian Process ($\mathcal{GP}$) prior [26]. This is referred to as a Gaussian Process Latent Variable Model (GP-LVM) [27]. Gaussian Processes can model any smooth mapping, and therefore represent a clear improvement over the linear mappings enforced by PCA.

A $\mathcal{GP}$ is a set of random variables, any subset of which follows a joint Gaussian distribution. The process is defined by a mean function $\mu(\cdot)$ and a co-variance function $k(\cdot, \cdot)$. In general we can, without loss of generality, center the data and set the mean function to be the constant $\mu(\mathbf{x}) = 0, \forall \mathbf{x}$. Regarding the covariance function (also referred as kernel function in the literature), it needs to generate a valid covariance matrix (positive semi-definite) when evaluated on any set of points in the input domain. In this paper we are going to take the standard approach and use an additive combination of a set of functions to parametrize the covariance of the $\mathcal{GP}$. In specific we are going to use a combination of a radial basis, bias and a white noise function,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_r\ e^{-\frac{\gamma}{2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)} + \sigma_b + \sigma_n \delta_{ij} \tag{7}$$

The radial basis $\sigma_r\ e^{-\frac{\gamma}{2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}$ governs how much different points in the dataset affect each other, based on their inter-distance. This term implicitly encodes a preference towards smooth function, since points close to other are more correlated than points far from each other, therefore generating similar observations. The bias term σ_b sets a minimum correlation between any pair of points. Finally the noise term $\sigma_n \delta_{ij}$ is used to "explain away" points that are not supported by the model (so that the covariance between some points is treated as noise).We will refer to the parameters of the covariance function $\theta = \{\sigma_r, \gamma, \sigma_b, \sigma_n\}$ as the hyper-parameters of the $\mathcal{GP}$. As we sill see in the next paragraph, these hyperparameters are optimized automatically so that they maximize the marginal likelihood of the data given the model.

Introducing a $\mathcal{GP}$-prior and combining it with the likelihood we can integrate out the generative mapping (i.e. considering all the possible mappings and integrating over them) leading to the marginal likelihood,

$$p(\mathbf{y}|\mathbf{X}, \sigma) = \int p(\mathbf{y}|\mathbf{f}, \sigma) p(\mathbf{f}|\mathbf{X}) \mathrm{d}f. \tag{8}$$

To remove the marginal likelihood's invariance to the scale of the latent space it is combined with an uninformative prior $p(\mathbf{X})$. The GP-LVM proceeds by finding the latent locations $\mathbf{X}$ and the hyper-parameters θ that maximize the marginal likelihood with respect to the observed data. Modeling each observed dimension with an independent $\mathcal{GP}$ leads to the following marginal likelihood,

$$P(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{j=1}^{D} \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} e^{-\frac{1}{2} y_j^T \mathbf{K}^{-1} y_j} \tag{9}$$

as a product of D independent Gaussian processes.

Finding the latent location and hyperparameters that minimize the negative logarithm of the marginal likelihood Eq. 9 is done by gradient based methods. For general covariance functions, the landscape of the objective function is likely to have several local minima which means we cannot guarantee that the global optima will be returned.

By using a smooth covariance function we encode a preference towards smooth generative mappings in the GP prior. This implies that points close in the latent space will remain close in the observed space. The opposite is not guaranteed though, i.e. points close in the observed space remain close in the latent space. However, this can be incorporated into the model by representing the latent locations $\mathbf{x}_i$ in terms of a smooth parametric mapping g_j from the observed data $\mathbf{y}_i$. In specific we are going to use a mapping that is capable of modelling non-linear correlations by employing a regression model over a kernel induced feature space,

$$x_{ij} = g_j(\mathbf{y}_i, a) = \sum_{n=1}^{N} a_{jn} k_{bc}(\mathbf{y}_i, \mathbf{y}_n), \tag{10}$$

where k_{bc} will be referred to as the back constraint kernel. This means that the maximum likelihood solution of the parameters a rather than the latent locations are sought.This is referred to as a back-constrained GP-LVM [28]. Practically, back-constraints force the existence of a functional mapping

between observed space $\mathbf{Y}$ and low dimensional space $\mathbf{X}$. This represents an efficient manner of projecting new observed data into an existing manifold $\mathbf{X}$.

The use of back-constraints represents a strong assumption which might significantly alter the solution. The kernel matrix of Eq. (10) is controlling the mapping in a similar way as in Equation 7. One difference is that the inverse width of the kernel is not optimized and has to be set, which gives some control over the latent trajectory smoothness. Having a very wide kernel means that the latent trajectories will be very smooth and this imposes a considerable constraint on the model which might reduce its ability to adapt to the data. On the other hand, if the kernel width is very narrow, the latent trajectories might become jagged. Additionally if one tries to project points that are far away (in terms of the kernel width) from the original dataset, the model might not have any evidence supporting this point to make a valid prediction. In that case the point will be projected to some point in the latent space where all unsupported points collapse to.

C. Spectral Non-linear Models

In addition to PCA there exists several non-linear spectral dimensionality-reduction models such as Isomap [29], MVU [30] and LLE [31]. Therefore it might seem unfair to compare the non-linear GP-LVM with the much more limited linear PCA. However, our reasons for focusing on the comparison with PCA are two-fold. First, PCA is the most commonly applied method in the field of synergy extraction, specially for grasping data, as we have discussed in Section II. Therefore, it is natural that we compare our proposed method with the state-of-the-art in the field. Second, our experiments confirm that the performance of LLE and Isomap on our data are worse than PCA, see Fig. 9 and Fig.18. It should be taken into account that these methods aim to find the intrinsic representation from local statistics in the data. Such statistics are much more uncertain and more severely affected by noise, reducing the applicability of such approaches.

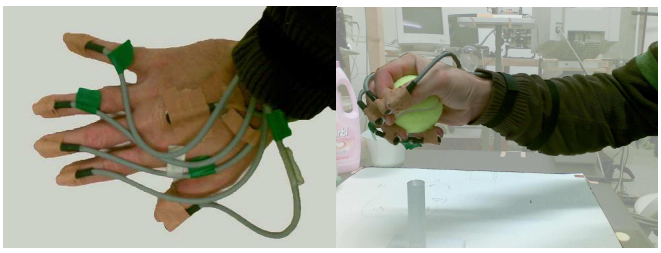
IV. Data Description

The extraction of postural synergies by exploiting dimensionality reduction techniques is based on the fundamental assumption that we can acquire a data-set which "well" describes the problems state domain *i.e.* being sufficiently densely sampled. In this section we will describe the data-set we created for the work in this paper. It is publicly available in http://grasp.xief.net/.

The data-set was generated from 5 subjects (3 male, 2 female). All subjects are right handed and have not reported any hand disabilities. The average hand length is 185.2 $\pm$ 13.3mm and hand width is 81.1 $\pm$ 7.4mm. A Polhemus Liberty system with six magnetic sensors was used for recording the data. Each sensor provided its orientation and position with respect to a base point as a 4d quaternion and a 3d vector (7 dimensions in total).

The spatial and angular resolution of each sensor is 0.8 mm and 0.15 degrees respectively. One sensor was applied to each fingertip, positioned on the fingernail and one was placed on the dorsum of the hand. See Fig. 5a for an image of the markers applied to the hand. The subjects were asked to perform 31 different grasp types [32] with their right hand on an object typical for the specific grasp. They were shown a picture of each grasp and a demonstration of the grasp was performed if the subject had problems mimicking the grasp on the picture. To start they placed the hand in front of them on the table in a flat hand posture. Upon a starting signal they grasped an object with the desired grasp type, lifted the object (this moment is shown in Fig. 5b), put it down again and retreated the hand to the starting position. The data recording started when the hand began to move and ended when the hand was put back to the initial position. As we are interested in studying the intrinsic posture of the hand we removed the global transformation of each grasp thereby representing each grasp in a common frame of reference.

Each grasp was discretized into 30 uniformly distributed time instances for which we recorded the fingertip poses. In summary, this means that we have acquired a database consisting of five subjects performing 31 different grasps resulting in 4650 datapoints in total. Furthermore, each subject was asked to perform each grasp twice. The first instance was used for testing and the second for training.



(a) Five sensors are placed on the fingertips and one on the wrist. (b) Grasp posture for grasp 11.

Fig. 5: Magnetic sensors setup

V. Synergetic Representations

In Fig. 6, a schematic figure of the evaluation framework is shown. We are particularly interested in evaluating the application of two different dimensionality-reduction approaches for extracting postural synergies. This is shown by the bottom-middle and bottom-right modules tagged as "PCA" and "GP-LVM" in Fig. 6. For representing temporal information as well as multiple subject variance in low dimensional space we use Gaussian Mixture Regression (GMR) [33] (bottom left module in Fig. 6). In this section, we will first briefly explain our usage of GMR (more information available in [33] and our previous work [34]) and then examine qualitatively the distribution of the low-dimensional representation extracted with different dimensionality reduction techniques.

A. Gaussian Mixture Regression of Grasps

Our modeling of low-dimensional action data has two main parts. First, a mixture of Gaussians is fitted to the data after extending it with a time dimension (if data was two-dimensional it becomes three-dimensional with a dimension

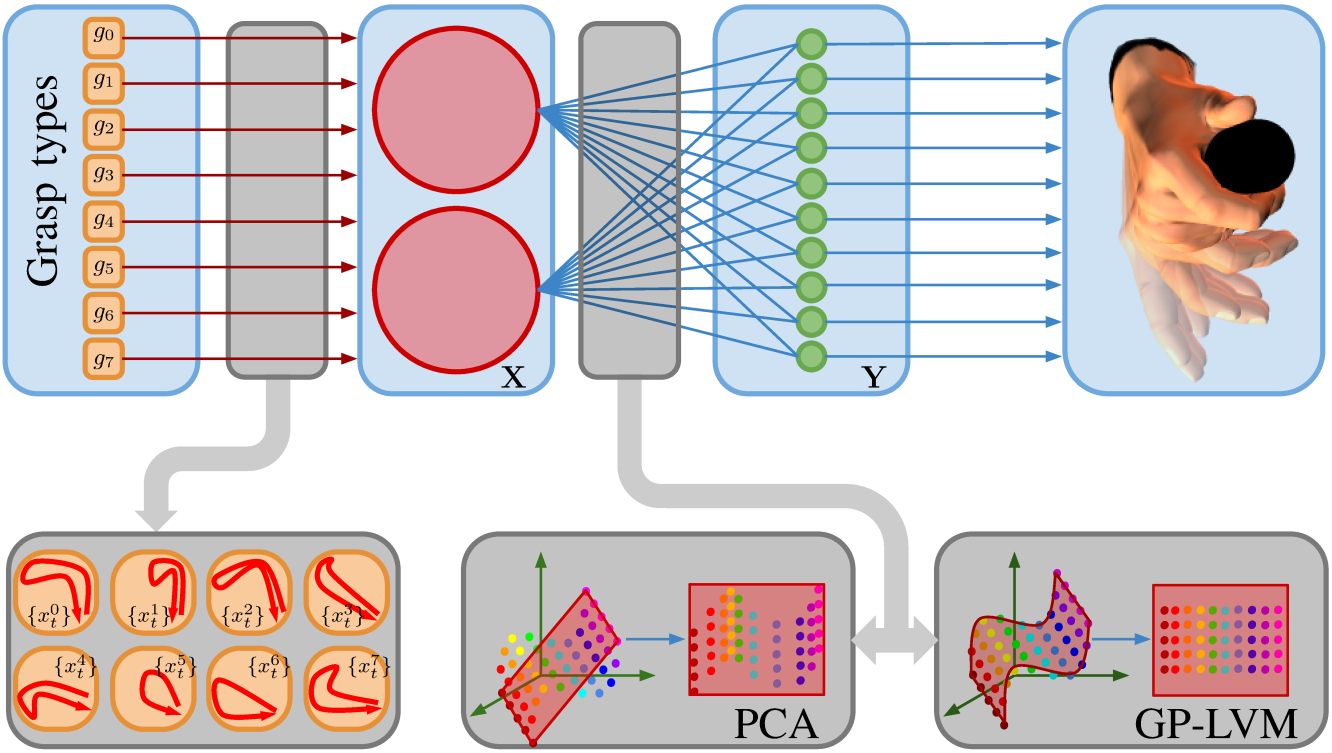


Fig. 6: The figure above shows the schematic description of the system for modeling grasping actions that we evaluate and analyse in this chapter. We model different grasp types g_i (top left box) as time series of poses $\{x_t^i\}$ in low dimensional space $\mathbf{X}$. We can see eight examples of those models in the lower left box. The low dimensional (two-dimensional in this example) space $\mathbf{X}$ is extracted from the high dimensional space $\mathbf{Y}$ using unsupervised dimensionality-reduction methods, such as PCA (bottom-middle) and GP-LVM (bottom-right). Processes. The hand poses (right) are fully described by the space $\mathbf{Y}$. The arrows in the figure follow the generative meaning of our model: grasp models describing low dimensional data that can be mapped into high dimensional space and be executed. However, the extraction process runs right to left: high dimensional executions mapped into low dimensional space that can be analyzed to create different grasp models.

representing time), second column of Fig. 7. This Gaussian Mixture Model (GMM) is computed by initializing the mixture of Gaussians with k-means and optimizing it through Expectation Maximization. Empirically, we found that using more than 3 Gaussians did not improve the generalization capabilities of the model, see Section VI-B2. Second, a hand posture is inferred for each time-step by using Gaussian Mixture Regression. This creates a continuous path through the latent space that describes the grasp (third column of Fig. 7, bottom left module of Fig. 6). That path has a mean and a variance. The paths corresponding to each of the 31 grasps can be found in Fig. 11.

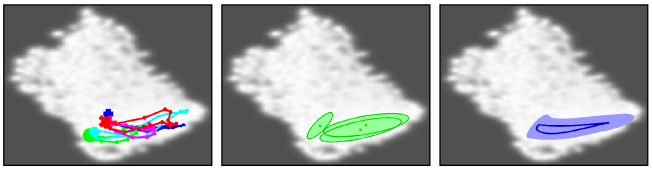


Fig. 7: Projection of grasp number 1 into latent space, GMM fitting, GMR regression. The other grasp types show similar patterns.

The GMM/GMR representation of the grasps is a powerful tool that can be used for several purposes. One is the generation of new actions under some constraints, [35]. In our case, this could help to generate an action composed of two grasps without coming back to the rest position between them. The second grasp can be constrained to start in a specific pose or after a specific time frame of the first grasp. The GMR can be optimized taking into account this constraint, providing thereby a smooth transition between those grasps.

In the remainder of this section we will present the representations extracted with PCA, GP-LVM, Isomap and LLE. First we plot the projection of the training data into the low dimensional representations. By examining how the trials of different subjects of the same grasp type group together and how the grasp types are distributed we can draw conclusions about how well the synergies represent the grasping data.

B. Principal Component Analysis

PCA finds and rotates the basis vectors of the space in such a way as that the first principal component (PC) explains most of the variance of the data. The second PC explains most of the remaining variance and so on. This reasoning brings up the natural question of how many dimensions are enough for properly representing the observed data. Fig. 8 shows how by increasing the number of principal components

the variance that is left unexplained decreases. The first component accounts for 59% of the variance, the second for 14% and the third for 5%. It is difficult to choose the dimensionality of the manifolds based on this data, since we have shown that even manifolds with high accuracy can result in grasp failure, Section II. Nonetheless, we decided to use two and three dimensional manifolds for three reasons: first, because their accuracy is similar to the accuracy reported in [7]; second, because the accuracy increases very slowly by adding further dimensions; and third, because the visualization of the manifolds is difficult for dimensionalities bigger than three.

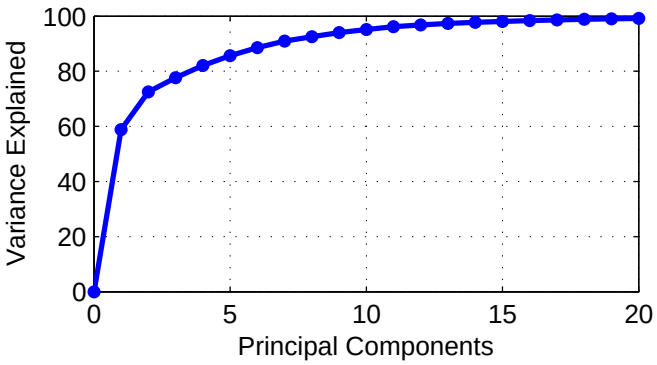


Fig. 8: Variance of the data (see Section IV) explained with increasing number of Principal Components. Only the first 20 components are shown, as the additional information transferred by the last ones is very small.

PCA 2D (Fig. 9a): To visualize the shape of the space, the datapoints were plotted as white dots over a dark background. The space is a rather narrow ark, on which all movements are situated. The initial starting posture is on the right side of the space. During the approach movement the trajectory progresses leftwards. The final position of the grasp is usually the point farthest away from the starting region. The overall flexure of the fingers determine how far the trajectory moves away from the start point. The reason for this is that the starting posture is a flat hand and therefore increasing the finger flexion increases the difference to the starting posture.

PCA 3D (Fig. 9b): The 3D plots add the third PC to the data. As expected, the variance is smaller than in the other two dimensions. The arc structure is still very dominant. The additional variance of the third PC is relatively small, nevertheless this additional information can be used to better distinguish between grasp types, as will be shown in Section VI-B.

C. Spectral non-linear methods

For the sake of conciseness we have not explained in depth the spectral non-linear methods. The reason is that their results are clearly worse than PCA and GP-LVM for our problem. Fig. 9c and 9f show that executions of the same grasp by different subjects are located in very different positions in the manifold and have very diverse directions (some trials in Fig. 9f are almost perpendicular to the rest). Moreover, since different trials of the same grasps are scattered around the manifold, different grasps will be hardly separable.

D. GP-LVM

Each point in latent space is connected via a Gaussian Process mapping to a point in high dimensional space. It predicts the mean and the variance of a point in high dimensional space given the latent location. The mean can be directly used as the reconstruction of a latent point in high dimensional space. The variance, which is connected to the prediction, can be used to quantify the confidence the model has while generating the point in high dimensional space. A large variance means that the model has a low confidence as it is poorly supported by data points. How fast the variance increases while moving away from data points gives a hint on the ability of the model to generalize to previously unseen points. For simplicity and coherence with the rest of the methods, Fig. 9 only shows the predicted mean. However, Fig. 7 shows the variance of the GP-LVM with the brightness of the background (white corresponding to low variance and dark corresponding to high variance).

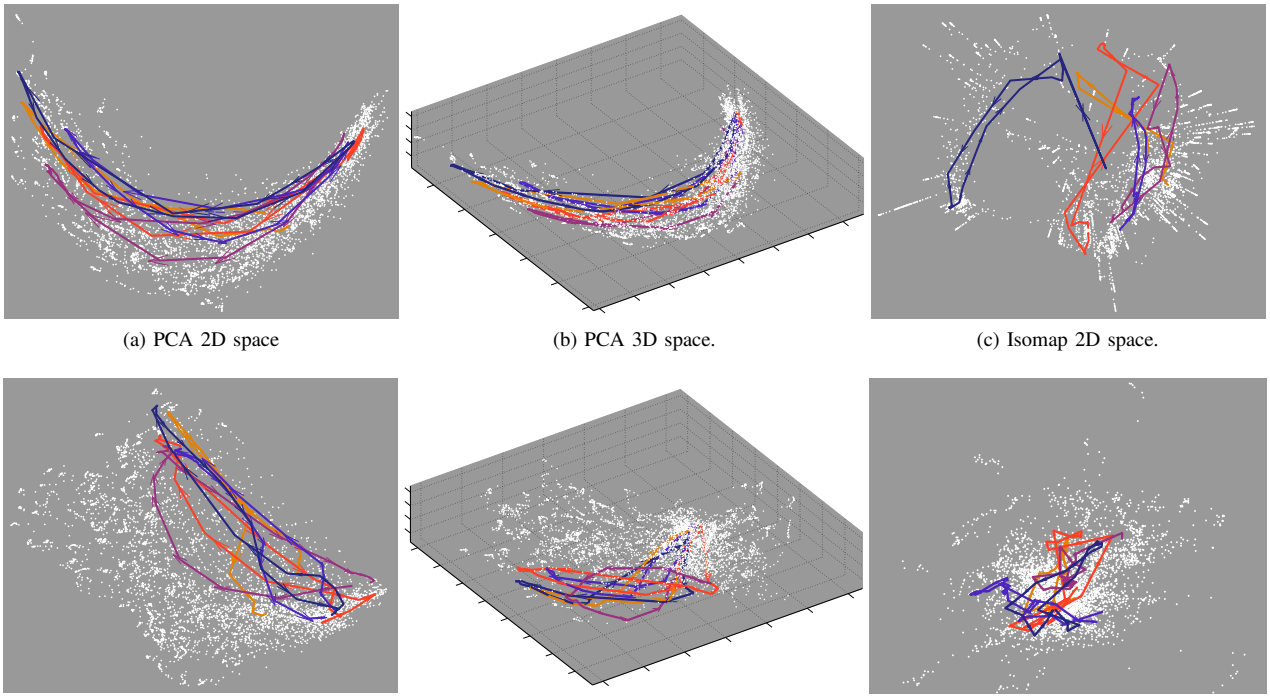
The GP-LVM 2D space (Fig. 9d) covers a larger area than the PCA 2D space (Fig. 9a). The non-linear character of GP-LVM allow it to spread the grasp types better and therefore having a finer differentiation between them, as we will further explore in Section VI-B. This comparison is valid as well for GP-LVM 3D and PCA 3D (Fig. 9e and Fig. 9b).

As we have seen in Section III-B, the kernel (or covariance) matrix is a core part of the GP-LVM methodology. In the following section we inspect the computed GP-LVM representation in terms of its kernel matrix.

1) Interpreting the Kernel Matrix: The information described in the previous section can be also visualized through the kernel matrix $\mathbf{K}_{i,i}$. In Fig. 10, we can see the kernel matrix corresponding to a 3D GP-LVM model, together with two magnifications of it. Dark tones correspond to low values (low correlation) while bright ones to high ones. Overall, the figure has 4650x4650 pixels, where the (i, j)-th pixel represents the correlation between the points i and j in latent space.

The data is composed of 31 blocks corresponding to different grasp types, each of them divided into 30 timestep blocks, and finally each timestep block is divided into 5 subjects. We can see that the highest correlation occurs around the diagonal, where 31 blocks are present. This means that points are highly correlated with points of the same grasp (even from different users or time instances).

If we observe the upper right magnification, we can see that the central square (grasping pose) that relates grasp 2 and grasp 3 has higher values than the one for grasp 1 and grasp 3. This means that the correlation between grasps 2 and 3 is higher than between grasps 1 and 3. If we observe the pictures corresponding to the grasps from Fig. 11, we can see that indeed the similarity between grasps 2 and 3 is higher than between grasps 1 and 3. These correlations can be observed in a better way if we crop the central part for each grasp, disregarding in this way the initial and final pose. Fig. 12a shows the kernel matrix with points corresponding to frames

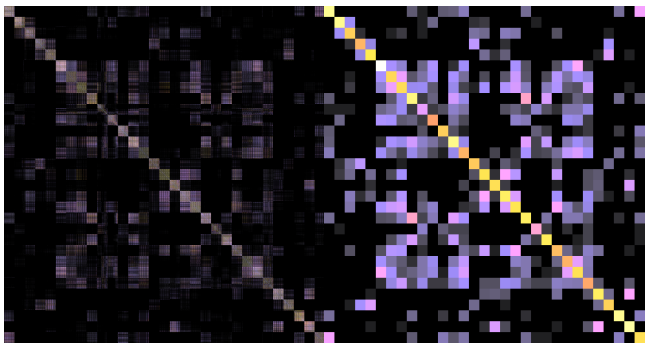


(a) PCA 2D space (b) PCA 3D space. (c) Isomap 2D space.

(d) GP-LVM 2D space. (e) GP-LVM 3D space. (f) LLE 2D space.

Fig. 9: Comparison of the synergetic representations of grasping data. In all figures the trajectories of the five subjects performing grasp number 4 are plotted in different colors. The white dots represent the projection of all the grasping points. Each axis represent one dimension in the low dimensional space.

10 to 20 (among 30 frames in total of one trial). Fig. 12b shows the sum of covariance values for the 10 time frames and 5 users inside of each grasp type. The values in this matrix can be interpreted as the covariance of the classes in the data, and can be used as a simple way of determining which grasp is similar to which one. This matrix is of size 31x31, as we have 31 different grasp types.



(a) Kernel matrix of grasping points (frames 10 to 20 in each grasp, approach and retreat removed).

(b) Mean covariance. Axes represent grasp types, and pixels represent the mean covariance of grasping datapoints (see left figure) in two grasps.

Fig. 12: Covariance matrix without approach and retreat datapoints.

In the lower right part of Fig. 10 we see the magnification of the covariance values corresponding to points of a single grasp. Along the diagonal we can identify three areas; the first and last are replicated along every row and column in the general kernel matrix, while the central one is only replicated for the grasps which are indeed similar to each other. That means that the first and last sections correspond to poses largely similar among users and grasps, i.e. the approaching and retreat phase of the grasp.

Another detail of the kernel matrix worth an explanation are the datapoints corresponding to grasp 21, as presented in Fig. 12b. Grasp 21 has the lowest mean correlation with itself among all the grasps. This is due to a large variance among subjects, as can be seen in Fig. 13a. It should be noted that the ordering in the left picture is that the first 5 pixels correspond to frame 1 of subjects 1-5, then frame 2 subjects 1-5 and so on. This creates this 5 pixel pattern. Once the matrix is rearranged to place the elements originating from the same user together (Fig. 13b), it becomes visible that subject 1 and 5 are completely uncorrelated from the rest, and correlation between subjects 2, 3 and 4 is low. We can interpret this as a sign of extreme variance in the execution of grasp 21. Indeed, our experience recording the grasp sequences was that this grasp was executed among the subjects in very different ways. Grasp 21 is the "Cigarette" grasp, where an object is placed between the index and the ring finger. We only demanded that the object is grasped properly, but we did not specify how the remaining fingers have to be positioned. Some subjects kept the remaining fingers extended, while others flexed them.

It should be mentioned that this analysis is not only applica-

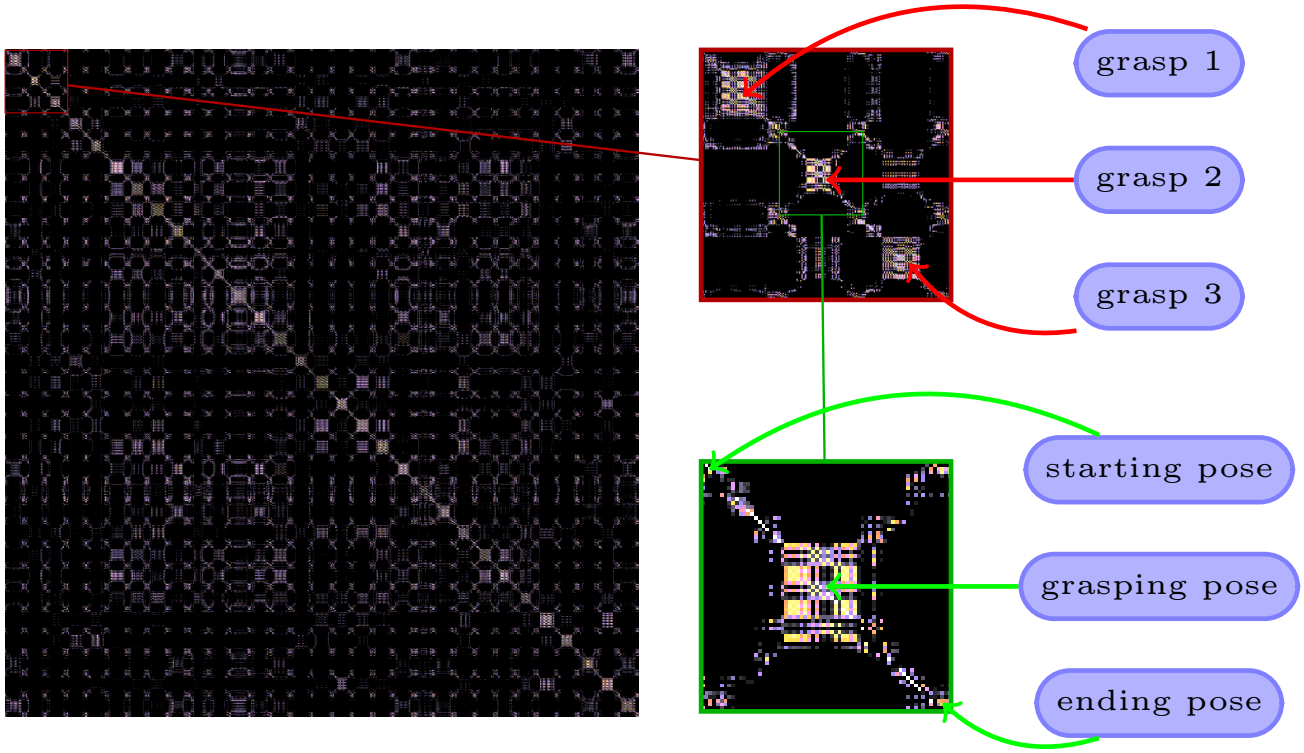
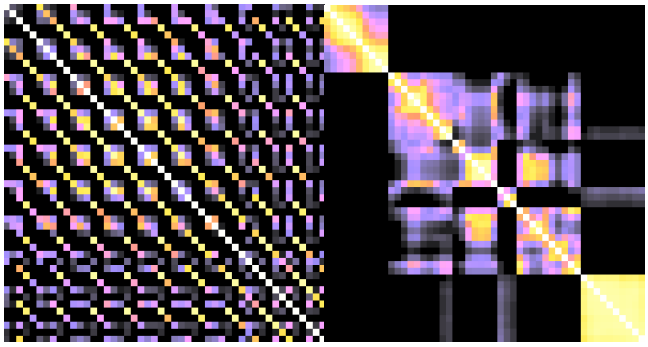


Fig. 10: Kernel matrix. Each axis represents the datapoints, and each (i, j)-th pixel the correlation between the datapoints i and j. Dark tones represent low correlations, whereas bright ones denote highly correlated points. The left picture is the full kernel matrix, whereas on the right two magnifications of the matrix can be seen. Relationships between datapoints such as similarities or motion structure can be directly assessed in the kernel matrix.



(a) Covariance grasp 21 (b) Covariance grasp 21 rearranged

Fig. 13: Correlation of poses tagged as grasp 21. In the right the elements are rearranged to group elements from the same user. This shows that different subjects performed grasp 21 in a very different way.

ble to grasping data, but to any data with dynamical behavior through time and multiple classes.

E. Summary

PCA and GP-LVM managed to extract a meaningful structure from the high dimensional data. In these models the trajectories of different subjects performing the same grasp type showed a similar pattern. Nevertheless, for a given dimension, GP-LVM is superior to PCA since it manages to spread the points over a larger area. This allows for a finer differentiation between grasp types while keeping different user instances of the same grasp close to each other, and therefore for a richer description. Both methods are able to generalize between subjects. That means that given one grasp type, the trajectories of the subjects all show a similar pattern and move along similar paths. That is the basis for the analysis in Section VI-B, where a model for each grasp type is generated using those five trajectories, defining in this way that particular grasp synergy.

VI. Evaluation

In this Section we will evaluate the grasp synergies proposed by GP-LVM, and compare them with the ones extracted using PCA. We evaluate in Section VI-A how well we can reconstruct poses that have been mapped to the low-dimensional space. Then, Section VI-B evaluates the semantics of the synergies, i.e. how compact are the models of each grasp and how well can we discriminate between them.

A. Evaluation of the Reconstruction Error

One important requirement on the extracted synergies is that they accurately represent the observed space of the data. Therefore, we evaluate the quality of the learned mapping

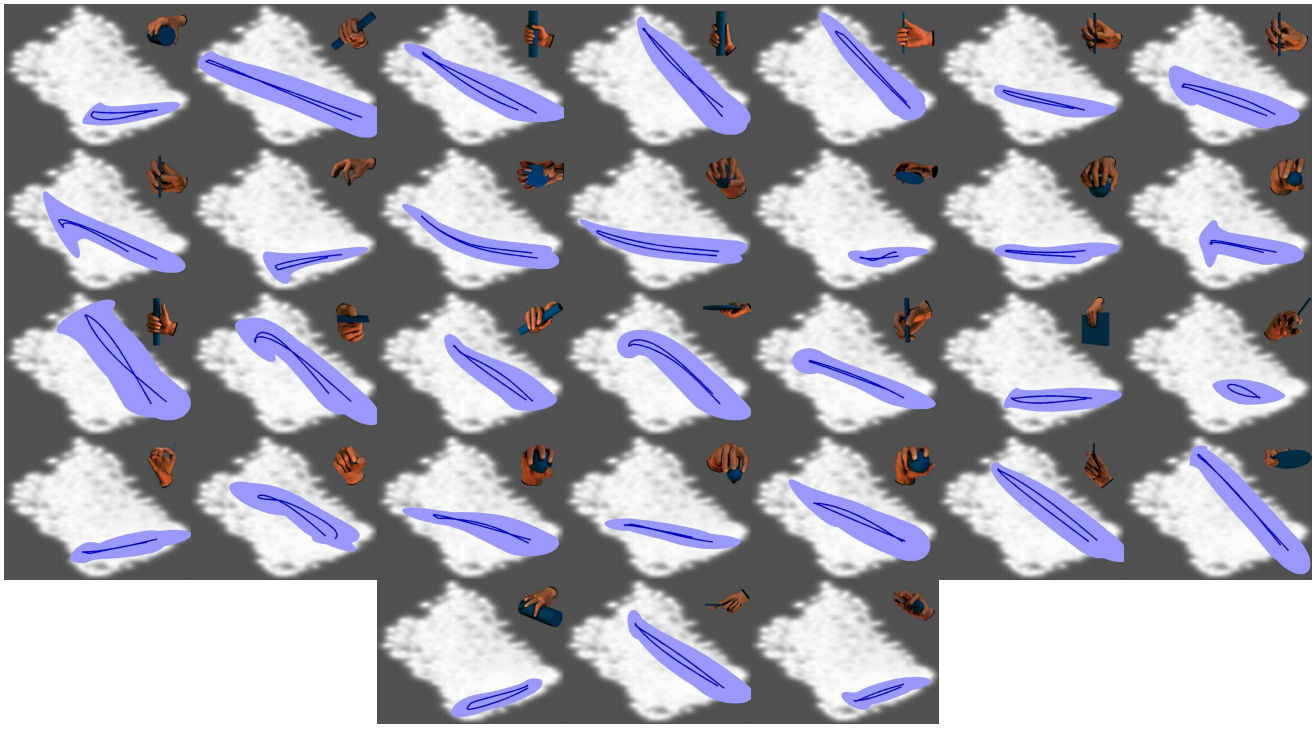


Fig. 11: GMR regression on the 31 grasp movements of all subjects. The dark line indicates the mean trajectory and the light area correspond to the uncertainty. The grasps are sorted, so the first row contains grasps 1 to 7 and so on.

in terms of the reconstruction error. The reconstruction error shows how much the mapping connecting the observed and the latent representation distorts the data. It is computed by pushing a point from the high-dimensional space through the latent-space and back to the original space. The reconstruction error is then the difference between the original and back-projected point. This is performed for the both the training points, which tests how well the model adapts to those points, and the test set, where it allows us to access the performance for points which are new, but similar to the training data. Since no information about the classes of the grasp types is included, it does not test the generalization ability of the model. The reconstruction error only tests how much information is lost in the mapping from high dimensional space to low dimensional space and back; it does not provide information about the semantics of the space, e.g. how similar the executions of a particular grasp by different subjects are.

For all four data sets the positional (Fig. 14) and rotational (Fig. 15) errors were calculated.

Any model created with GP-LVM outperforms all models created by PCA in terms of reconstruction error. It is worth consideration the difference in performance between training and test data. In both GP-LVM models the training data has lower errors (both positional and rotational) compared to the test data. Interestingly such a trend is not visible for PCA where the error on training and test data are very similar. This is due to the fact that the synergies from PCA tend to be over-smoothed, average trajectories (see Fig. 16). Such average trajectories are “equally wrong” for training and testing. GP-LVM adapts better to the trajectories at the cost of a slight overfit. Nevertheless, the reconstruction error of GP-LVM is around 20% better than the error from PCA of the same dimensionality.

Increasing the dimensionality of the latent space allows to better fit the training data onto the manifold, decreasing the reconstruction error for the training set. Similar effects are observed for the test set. suggesting that the higher dimensional models are not overfitting the data. Importantly though, the decrease is significantly larger for the non-linear GP-LVM indicating that the correlations in the observed space are non-linear and cannot be modeled using PCA.

Both algorithms seem to treat positions and rotations with equal importance and in a similar fashion. If one compares the figures of positional (Fig. 14) and rotational errors (Fig. 15) the relative differences between models and test/training set are very similar.

The overall performance of the extracted synergies by GP-LVM are better compared to the PCA synergies of the same dimensionality, as it has better results for the training set as well as for the test set. Human hand motion in general is non-linear and therefore an algorithm that can cope with non-linearities (such as GP-LVM) in the data will be superior. Additionally PCA looks for the largest variance in the data, which might be dominated by noise and the valuable information is blurred. GP-LVM, being a probabilistic approach, also has the ability to explicitly model the noise which increases the performance of the algorithm.

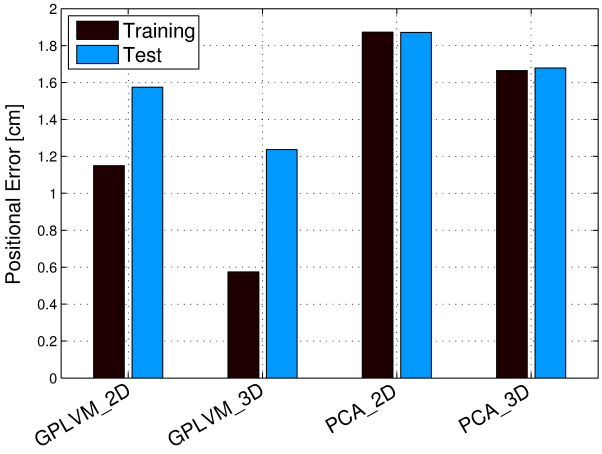


Fig. 14: Reconstruction errors regarding the position of the training and test data sets.

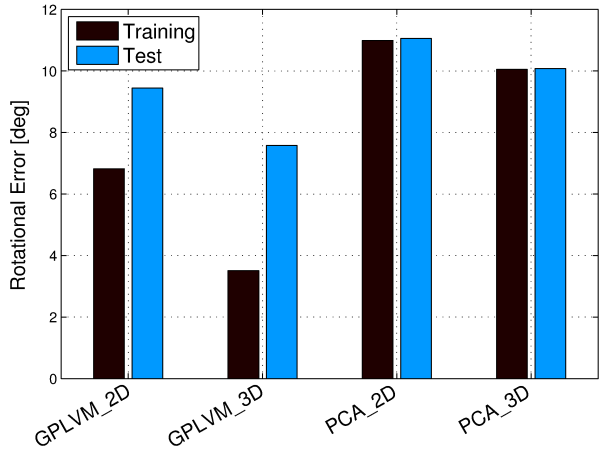


Fig. 15: Reconstruction errors regarding the orientation of the training and test data sets.

B. *Semantic Evaluation*

Low reconstruction error is not the only desirable characteristic. The representation should also capture and reflect semantic details of the data, like similarity among different users executing the same grasp or dissimilarity among different grasps.

1) Visualization of the mean grasp model: To assess how well the GMM/GMR models fit to the original data, we project the latent trajectory of the GMM/GMR model back to the high dimensional space. The comparison to the original data gives insight into how good the created GMM/GMR grasp models are. We reduced the amount of data displayed for the sake of clarity. The movements are projected to the plane spanned by the palmar-distal directions. Fig. 16 shows the fingertip movements of the index finger and the thumb projected onto that plane. The top image shows the corresponding grasp type as well as the plane the movements are projected onto. In the background of the other images the original movements of the

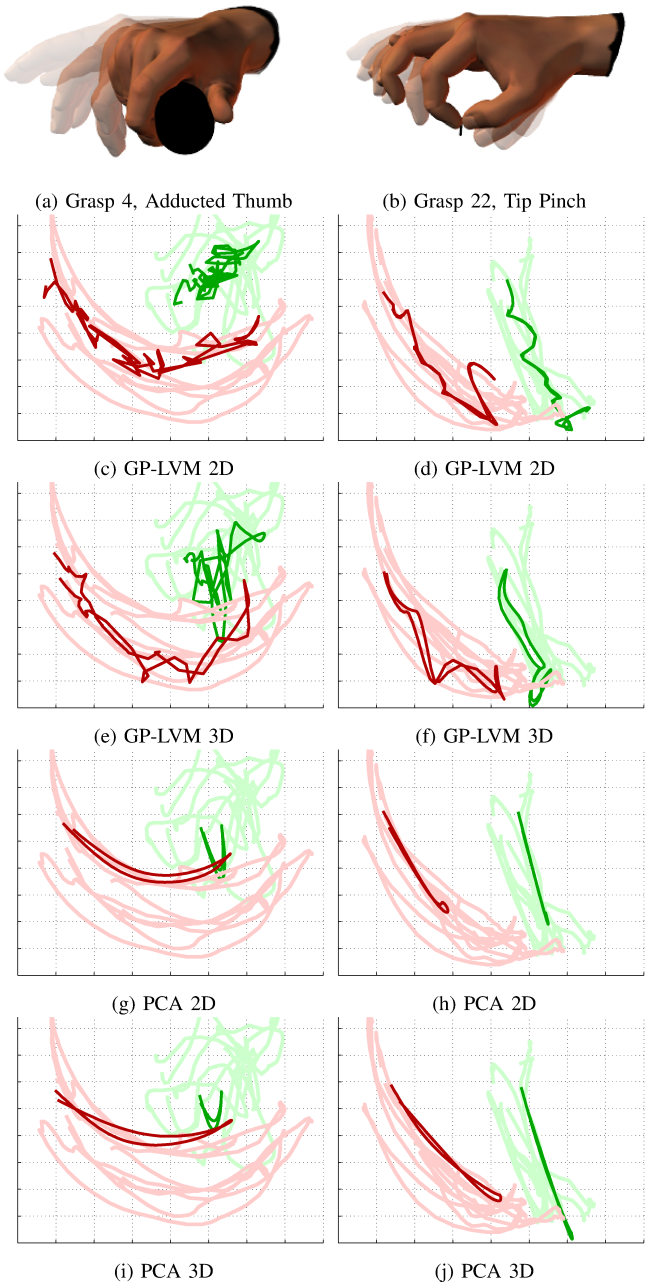


Fig. 16: Projection of two grasp types, as shown in the top figure. The second row shows the GMM/GMR model of that grasp type. Both grasp types show a distinct pattern in latent space, and the trajectory is smooth. The next four rows represent the projection of the latent space trajectory back to the high dimensional space. The spaces are, from top to bottom: GP-LVM 2D, GP-LVM 3D, PCA 2D, PCA 3D. The right trajectories correspond to the index finger and the green trajectories correspond to the thumb. The lighter trajectories in the background correspond to the original five executions of the grasp, while the highlighted one represents the computed model.

5 subjects are shown in a lighter color.

The grasp on the left side is a special variation of the power grasp, where the thumb is aligned with the axis of the cylinder. In this grasp type the thumb is relatively static, as it only abducts for the grasp. Since abduction/adduction is a movement largely perpendicular to the plane, most of the movement is lost by the projection. This makes the appearance of the thumb relatively random, but the 3D trajectory shows a distinct pattern.

The grasp on the right side is a precision grasp, where the index and the thumb are used to pick up a small object. Therefore, it is important that those two digits are close in the actual grasping phase. The background trajectories in Fig. 16 clearly show this.

GP-LVM produces relatively rugged trajectories, but they follow the subjects trajectories quite well. They have roughly the same range of motion as the subjects. In the precision grasp (right column) they reach a position where thumb and index finger are very close, which is a functional requirement of the grasp type. The three dimensional GP-LVM is smoother and it's trajectories fit the original ones even better.

PCA produces very smooth curves, but it cannot create the curved path of the subjects trajectories. There is an offset and the trajectory cannot follow the full motion amplitude of the subjects. When the dimension is increased to three, the shape of the trajectory improves – the curvature gets a little bit larger and the length of the trajectory better fits the subject's one.

Overall GP-LVM outperforms PCA, since it is able to follow the path of the human fingertips much better for a given dimension. That comes at the cost of having more ragged trajectories. In most applications this is more desirable than having smooth trajectories, which follow the wrong path.

The rotational component of the fingertip cannot be easily visualized, so a comparable analysis on the rotations was not performed. Nevertheless it can be assumed that they will behave in a very similar fashion, as the reconstruction error is very similar in positions and orientation (Section VI-A).

2) Grasp similarity: In this section we classify grasping actions according to the GMM/GMR models, see Section V-A. In Fig. 11 the evaluation of the trajectories from the dynamical model applied to the GP-LVM 2D representation is shown. Following the process depicted by Fig. 7, each grasp model is based on five trajectories, as performed by the subjects. The dark line corresponds to the mean trajectory and the light area shows the variance the model has on certain points of the trajectory. One can clearly observe that different trajectories have a different signature in the latent space.

Computing the similarity between grasps (pose sequences) is not straightforward. We will use the probabilistic description of grasps in terms of Gaussian Mixture Models for this purpose. Based on these probabilistic models, we can compute how likely it is that each point x in the space is generated by a grasp g_i.

$$p(x|g_i) = \sum_{k=1}^{K} \pi_k^{g_i} \mathcal{N}(x|\mu_k^{g_i} \sigma_k^{g_i}) \tag{11}$$

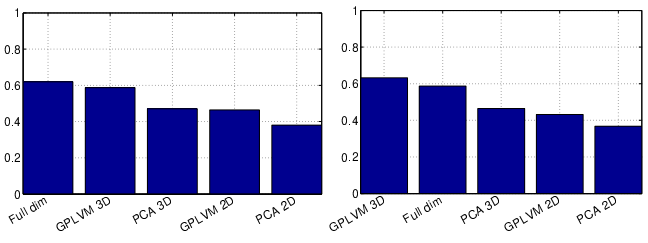
$$p(g_j|g_i) = \prod_{\forall x \in g_i} p(x|g_i) \tag{12}$$

$$s(g_j, g_i) = (p(g_j|g_i) + p(g_i|g_j))/2 \tag{13}$$

Equation 11 states that the probability of a point x belonging to a grasp g_i is modeled as a weighted mixture of gaussians, as explained in Section V-A, once we make the simplifying assumption of independence between the poses, Eq. 12. Other methods like HMM matching of sequences could be applied here instead, [36].

Note that this probability is not symmetric: $p(g_j|g_i) \neq p(g_i|g_j)$. We define the similarity between two grasps $s(g_j, g_i)$ as the average of those two quantities, Eq. 13.

Following these equations we can compute the probability of a new grasp sequence having been generated by a particular GMR model. By comparing those probabilities we can estimate which is the most likely grasp class that generated that sequence, and compare it with the real grasp that was actually executed. We performed this classification task for the grasp actions in the test set (not used for training) In Fig. 17 we can see that GPLVM 3D manages to generalize the model over new sequences equally well or better than the full-dimensional representation, while using only three dimensions instead of 35. GP-LVM is consistently better than PCA for a given manifold dimensionality. The amount of Gaussians used in the GMM/GMR model does not make large differences in performance.

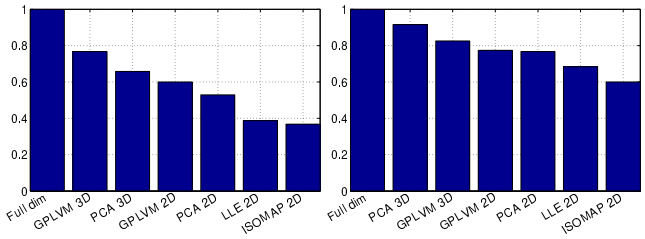


(a) 3 Gaussians (b) 6 Gaussians

Fig. 17: Classification rate for GMM/GMR models. GP-LVM outperforms PCA for a given dimensionality, and performs similarly to the full dimensional model, which uses more than 10 times more dimensions

It is not possible to perform this classification task with Isomap and LLE manifolds, since the standard version of these methods do not provide a way of projecting new data not existing in the training set onto the lower dimensional manifold. Therefore, for LLE and Isomap we can only show how well they separate the training data in lower dimensional space, by classifying the training set based on the models extracted from the same set. Fig. 18 shows the classification performance for the methods in Fig. 17 plus Isomap 2D (Fig. 9c) and LLE 2D (Fig. 9f). The first observation we can make is that the models computed in full-dimensional space

perfectly separate the data. However, we should remember that those models generalize over new data similarly or worse than the GPLVM 3D models. Second, we can observe that PCA performs well in this classification task; this tells us that it clearly overfits the training data, since its performance is much worse in Fig. 17. The fact that the classification capabilities of mixtures of 6 Gaussians is only better for training data indicates that mixtures of more than 3 Gaussians produce overfitting. LLE and Isomap perform clearly worse than the two-dimensional versions of PCA and GP-LVM, as we expected from the shape of their manifolds in Fig. 9. We should remember however that this classification is much more sensitive to the particularities of the training set than the classification used in Fig. 17.



(a) 3 Gaussians (b) 6 Gaussians

Fig. 18: Classification rate for GMM/GMR models tested with the training data itself. Although the grasp training set is more separable in PCA 3D and full dimensionality than in GPLVM 3D for 6 Gaussians, the latter generalizes better over new data and therefore outperforms the rest when classifying previously unseen data (Fig. 17b). Isomap and LLE clearly perform worse than the rest of the methods.

VII. Conclusions

The work presented in this paper relates to two important areas of robotics: i) human observation and motion analysis, and ii) representations that enable successful action planning and control. In applications that consider hand activities, a common way of controlling grasping actions for robotic hands is to use high-dimensional human grasping data. Different representations based on dimensionality reduction techniques have been used to enable viable planning and control solutions. Commonly, postural synergies have been used as a low-dimensional representation to enable correspondence between human and robot hand activities. The technique was developed to increase the understanding of complex relationships between the joints and muscles in human hands. As such, the original work on postural synergies was based on linear dimensionality reduction methods which, as we have shown in this paper, do not represent the human hand activity in an appropriate manner due to the inherent non-linearities in the data. We have argued that this significantly limits the usefulness of postural synergies as a modelling paradigm and that non-linear dimensionality techniques should be exploited to represent the data in a more appropriate manner.

The work presented in this paper addressed the non-linear dimensionality-reduction methods and their application to encoding highly non-linear human grasping data. Apart from encoding of postural synergies, our work relates closely to recent work in control of combined reaching and grasping movements in robots. However this work is built on assumptions of a causal relationship between reaching and grasping, something that may not hold. An illustrative example in the beginning of the paper and detailed discussion of related work serve as a motivating example of the applicability of the proposed approach.

As the first contribution, we have shown that non-linear dimensionality reduction methods can be used to tackle the correlations problem without considering causal relations between dimensions, but by assuming them to have been generated from an external manifold which we infer from data. Our second contribution is a thorough analysis of the internal parameters used in dimensionality-reduction techniques, sheding light into algorithms which have been traditionally used as a "black-box". Finally, we have provided an extensive experimental evaluation that showed how the proposed methods outperform the standard techniques in the field both in terms of recognition and generation of motion patterns. To this end, we have presented both qualitative and quantitative results of applying two different approaches for learning low-dimensional representations of hand pose data.

References

[1] Y. Demiris and G. Hayes, *Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model*. Cambridge, MA, USA: MIT Press, 2002, pp. 327–362.

[2] M. Ciocarlie and P. Allen, "Data-driven optimization for underactuated robotic hands," in *2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*. IEEE, May 2010, pp. 1292–1299.

[3] A. Shukla and A. Billard, "Coupled dynamical system based armhand grasping model for learning fast adaptation strategies," *Robotics and Autonomous Systems*, no. 0, pp. –, 2011.

[4] C. Granville, D. Southerland, J. Platt, and A. H. Fagg, "Grasping affordances: Learning to connect vision to hand action," pp. 59–80, 2009.

[5] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008.

[6] L. H. Ting, "Dimensional reduction in sensorimotor systems: a framework for understanding muscle coordination of posture." *Progress in Brain Research*, vol. 165, pp. 299–321, 2007.

[7] M. Santello, M. Flanders, and J. Soechting, "Postural hand synergies for tool use," in *The Journal of Neuroscience*, 1998.

[8] C. R. Mason, J. E. Gomez, and T. J. Ebner, "Hand synergies during reach-to-grasp," *J Neurophysiol*, vol. 86, no. 6, pp. 2896–2910, December 2001.

[9] I. V. Grinyagin, E. V. Biryukova, and M. A. Maier, "Kinematic and dynamic synergies of human precision-grip movements," *Journal of Neurophysiology*, vol. 94, no. 4, pp. 2284–2294, 2005.

[10] F. J. Valero-Cuevas, F. E. Zajac, and C. G. Burgar, "Large index-fingertip forces are produced by subject-independent patterns of muscle excitation," vol. 31, pp. 693–703+, 1998.

[11] M. Gabiccini, A. Bicchi, D. Prattichizzo, and M. Malvezzi, "On the role of hand synergies in the optimal choice of grasping forces," *Auton. Robots*, vol. 31, no. 2-3, pp. 235–252, 2011.

[12] M. C. Tresch, V. C. K. Cheung, and A. d'Avella, "Matrix factorization algorithms for the identification of muscle synergies: Evaluation on simulated and experimental data sets," *Journal of Neurophysiology*, vol. 95, no. 4, pp. 2199–2212, 2006.

[13] M. T. Ciocarlie and P. K. Allen, "Hand posture subspaces for dexterous robotic grasping," *I. J. Robotic Res*, vol. 28, no. 7, pp. 851–867, 2009.

[14] C. L. Nehaniv and K. Dautenhahn, *The correspondence problem*. Cambridge, MA, USA: MIT Press, 2002, pp. 41–61.

[15] S. B. Kang and K. Ikeuchi, "Toward automatic robot instruction from perception: Mapping human grasps to manipulator grasps," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 1, pp. 81–95, 1997.

[16] E. W. Weisstein. (2011, Dec.) Function. from mathworld–a wolfram web resource. [Online]. Available: http://mathworld.wolfram.com/Function.html

[17] I. T. Jolliffe, *Principal Components Analysis*. Springer-Verlag, 1986.

[18] A. Tsoli and O. C. Jenkins, "Neighborhood denoising for learning high-dimensional grasping manifolds," in *IROS*, 2008, pp. 3680–3685.

[19] K. Nguyen and V. Perdereau, "Arm-hand movement: Imitation of human natural gestures with tenodesis effect," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, sept. 2011, pp. 1459 –1464.

[20] J. Steffen, M. Pardowitz, and H. Ritter, "A Manifold Representation as Common Basis for Action Production and Recognition," *KI 2009: Advances in Artificial ...*, 2009.

[21] S. Bitzer and S. Vijayakumar, "Latent Spaces for Dynamic Movement Primitives," *International Conference on Humanoid Robots*, 2009.

[22] D. Pratichizzo, M. Malvezzi, and A. Bicchi, "On motion and force controllability of grasping hands with postural synergies," in *Proceedings of Robotics: Science and Systems*, Zaragoza, Spain, June 2010.

[23] M. Malhotra and Y. Nakamura, "The relationship between actuator reduction and controllability for a robotic hand," in *IEEE International conference on Biomedical Robotics and Biomechatronics*, 2010.

[24] R. Diankov, "Automated construction of robotic manipulation programs," Ph.D. dissertation, Carnegie Mellon University, Robotics Institute, August 2010. [Online]. Available: http://www.programmingvision.com/rosen_diankov_thesis.pdf

[25] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[26] C. Rasmussen and C. Williams, "Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)," 2005.

[27] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *The Journal of Machine Learning Research*, 2005.

[28] N. D. Lawrence and J. Quinonero-Candela, "Local distance preservation in the gp-lvm through back constraints," in *ICML06*, pp. 513–520.

[29] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, 2000.

[30] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *International Conference on Machine Learning*, 2004.

[31] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 2000.

[32] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, June 2009.

[33] S. Calinon, *Robot Programming by Demonstration: A Probabilistic Approach*. EPFL/CRC Press, 2009.

[34] J. Romero, T. Feix, H. Kjellström, and D. Kragic, "Spatio-temporal modeling of grasping actions," in *IROS*. IEEE, 2010.

[35] S. Calinon, F. Guenter, and A. Billard, "On learning, representing and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 37, pp. 286–298, 2007.

[36] S. Calinon and A. Billard, "Recognition and reproduction of gestures using a probabilistic framework combining pca, ica and hmm," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 105–112.