2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

ST-HMP: Unsupervised Spatio-Temporal Feature Learning for Tactile Data

Marianna Madry Liefeng Bo Danica Kragic Dieter Fox

***Abstract*— Tactile sensing plays an important role in robot grasping and object recognition. In this work, we propose a new descriptor named Spatio-Temporal Hierarchical Matching Pursuit (ST-HMP) that captures properties of a time series of tactile sensor measurements. It is based on the concept of unsupervised hierarchical feature learning realized using sparse coding. The ST-HMP extracts rich spatio-temporal structures from raw tactile data without the need to pre-define discriminative data characteristics. We apply it to two different applications: (1) grasp stability assessment and (2) object instance recognition, presenting its universal properties. An extensive evaluation on several synthetic and real datasets collected using the Schunk Dexterous, Schunk Parallel and iCub hands shows that our approach outperforms previously published results by a large margin.**

I. Introduction

The rapidly advancing tactile sensing technologies provide sensory data of increasing quality and are commonly used in robotics applications. The interplay between visual and tactile sensing is paramount for the interaction of the robot with the real world: while visual input serves as a natural source of information for scene understanding, segmentation and object detection [1]–[5], tactile information is crucial for object grasping and manipulation.

Tactile sensing can be used to localize an object in a robot's hand and determine its material and shape properties [6], [7]. It can be used to estimate grasp stability [8], [9] and allow for re-grasping if slippage occurs [10]. Furthermore, tactile sensing is a valuable source of information for object instance recognition [11]–[13], especially when a part of the object is visually occluded. Finally, the relation between an object type and tactile sensing has been exploited to ensure that the object affords the assigned task and the robot manipulates it in a suitable way (e.g. a bottle is grasped so that pouring can be achieved [14], [15]).

Those examples demonstrate that an ideal representation of tactile data should serve a wide variety of applications and have the capability to adapt to the specific requirements of applications. This may be achieved using unsupervised feature learning techniques [16], [17] or deep learning methods [18]. Often, representations are obtained by manually specifying geometric properties of object imprints on the sensor [12], [19]. In contrast, feature learning results in representations generated from "raw" input signals without the need of specifying discriminative characteristics a-priori.

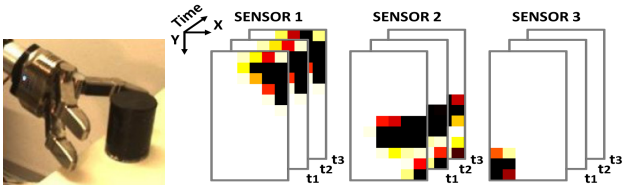


Fig. 1. Example of a series of tactile data obtained for the three finger Schunk Dextrous Hand (SDH) [21]. The measurements create a sequence over time in which information in consecutive frames is strongly correlated.

An optimal representation should also encode all the relevant dependencies existing in the data. As illustrated in Fig. 1, tactile data readings create a sequence in which consecutive frames are strongly correlated over time. Thus, as demonstrated in our experiments, exploiting the temporal dimension becomes particularly important in the tactile domain.

A. Approach and Contributions

We propose a representation for tactile data based on an unsupervised feature learning approach, the Hierarchical Matching Pursuit (HMP) [20]. The HMP was shown to outperform not only traditional descriptors such as SIFT, but also kernel-based feature learning methods and convolutional deep belief networks [20]. The HMP was originally designed for image classification and, as such, represents each spatial sample (image) separately. It builds multi-layer, rich feature representation by using sparse coding techniques and *spatial* pooling over image dimensions.

In this work, we extend that framework to *spatio-temporal* feature learning. The main idea is to extract features from consecutive frames and then pool them over the time dimension. This process is repeated at several scales of a spatio-temporal pyramid to capture all the relevant characteristics.

We show effectiveness of our approach on two examples of popular tactile-based robotics applications: grasp stability assessment and object instance recognition. Our Spatio-Temporal HMP yields excellent results and outperforms the state of the art for several synthetic and real databases that were collected using several different grippers, such as the Schunk Dextrous, Schunk Parallel and iCub hands. We demonstrate that our framework can be directly applied to data with diverse characteristics without any change in its design. We believe that the presented solution can be useful for other applications employing tactile data and to represent other spatio-temporal information, such as videos.

M. Madry and D. Kragic are with the Centre for Autonomous Systems and the Computer Vision & Active Perception Lab, CSC, KTH Royal Institute of Technology, Stockholm, Sweden. {madry,danik}@csc.kth.se Liefeng Bo is with Amazon Inc. The work was completed while he was affiliated with the Intel Science and Technology Center on Pervasive Computing, Seattle, USA. {liefengbo@gmail.com}. Dieter Fox is with the Department of Computer Science & Engineering, University of Washington, Seattle, USA. {foxg@cs.washington.edu}.

II. Related Work

A common approach to representing tactile data is to manually craft the representation based on prior knowledge about the properties of the inputs. Early works aimed at identifying simple primitive shapes in pressure patterns, such as points or lines [7], [22]. Since, this approach was shown to be inefficient, other methods focused on higher level geometric properties of object imprints in the tactile matrices. After identifying patches in the sensor readings that are likely to be contact regions, those methods describe patch characteristics using a set of manually defined geometric attributes: position, area and eccentricity of a patch [12] or higher-order moments [19], [21], [23]. Often, such statistics are combined together and as feature vectors provided directly to a classifier [21]. However, as shown in [19], some of these statistics might not be relevant for the task and dimensionality reduction have been applied to retain the most relevant properties [12], [24]. Other approaches draw from Computer Vision techniques. In [25] authors adapted image descriptors such as different filter sets and SIFT.

In terms of the key idea of obtaining features directly from raw data, the most related to our approach are [11], [26], [27], and [28]. However, none of those uses the state-of-the-art unsupervised feature learning methods. In [28] authors apply covariance analysis to sensor data to identify basic local object shapes. In [26] input measurements are transformed to binary matrices small fragments of which are directly represented as codes. However, to compute the final histogram representation all combinatorial possibilities are used and no actual feature learning takes place. In [27], although authors learn the representation from data, the method is limited to the application of shape-based object recognition. The object model consists of a mosaic of geometric patches aligned according to sensor positions. In [11] the bag-of-words approach is used and the dictionary is learned from low resolution sensor input using the K-means algorithm.

Intuitively, including temporal information to the representation should be beneficial. While a hand is closing around an object, it touches different parts of the object and applies different force, see Fig. 4 (Bottom). However, several approaches to analyzing grasp stability and object recognition use only on a single reading acquired at the end of the process ignoring temporal properties [9], [11], [28]. A less common approach is to model series of data. In [21] in order to assess grasp stability, sequences were represented by the Hidden Markov Models (HMM). Series of tactile data were also used to compare a newly proposed type of tactile sensors with previous hardware in [23]. Authors applied dynamic programming techniques to classify data collected for different household objects. An interesting solution has been proposed in [29] where a Gaussian Process with a recursive kernel has been used for object recognition. However, in all these studies *pre-defined* types of features have been employed, such as higher order moments or geometrical properties of contact regions, as listed in Tab. 2. We will compare our method with these approaches in Section IV-C.

III. Spatio-temporal Hierarchical Matching Pursuit

In this work, we propose the *spatio-temporal Hierarchical Matching Pursuit (ST-HMP)* that builds on the recently introduced Hierarchical Matching Pursuit (HMP) [20] algorithm. The HMP is a multilayer sparse coding network that creates feature hierarchies from raw data, layer by layer, with an increasing receptive field size. This approach has been successfully used for object recognition and achieved superior performance on standard vision recognition benchmarks [17]. However, the original HMP is limited to spatial signals such as images and depth maps.

In grasping tasks considered here, temporal information is critical for good accuracy. The ST-HMP extracts rich temporal structures from raw tactile data without pre-defining discriminative data characteristics. It uses K-SVD to learn codebooks in an unsupervised fashion over a large collection of spatial (or spatio-temporal) patches sampled from tactile data. With the learned codebooks, the ST-HMP computes sparse code sequences using orthogonal matching pursuit and then pools them in a *spatio-temporal* pyramid manner to achieve robustness to both spatial and temporal variations.

A. Dictionary learning

We use K-SVD [30], a popular codebook learning approach to learn the codebooks from tactile data. The key idea of K-SVD is to represent data as sparse linear combinations of codewords selected from a codebook and has achieved the state-of-the-art results in various low-level image processing tasks such as image denoising and image compression. K-SVD learns the codebook $D = [d_1, \cdots, d_m, \cdots, d_M] \in R^{H \times M}$ and the associated sparse codes $X = [x_1, \cdots, x_n, \cdots, x_N] \in R^{M \times N}$ from a matrix $Y = [y_1, \cdots, y_n, \cdots, y_N] \in R^{H \times N}$ of observed data by minimizing the reconstruction error

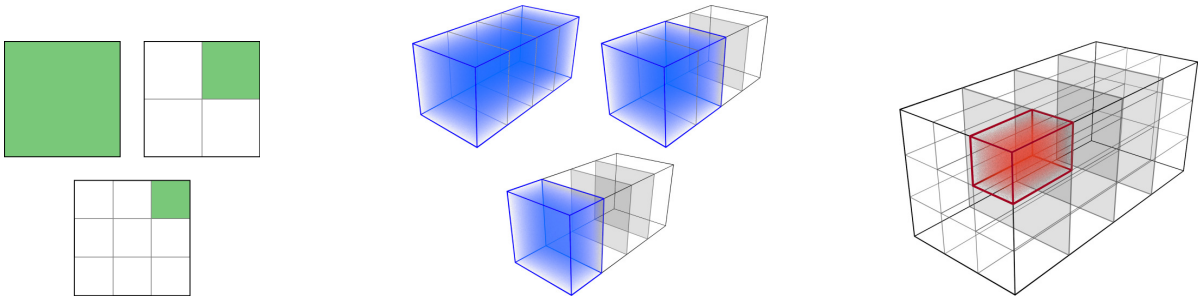
$$\min_{D,X} \|Y - DX\|_F^2 \qquad (1)$$
$$s.t.\ \forall m,\ \ \|d_m\|_2 = 1 \ \text{ and } \ \forall n,\ \ \|x_n\|_0 \leq K$$

where H, M, and N are the dimensionality of the codewords, the size of the codebook, and the number of training samples, respectively, $\|\cdot\|_F$ denotes the Frobenius norm, the zero-norm $\|\cdot\|_0$ counts non-zero entries in the sparse codes x_n, and K is the sparsity level controlling the number of the non-zero entries. When K-SVD is applied to our grasping tasks, the data matrix Y consists of raw spatial patches randomly sampled from tactile sequences.

K-SVD solves the optimization problem 1 in an alternating manner. During each iteration, the current codebook D is used to encode the data Y by solving an inference problem

$$\min_x \|y - Dx\|^2 \qquad s.t.\ \|x\|_0 \leq K \qquad (2)$$

This problem is also known as compressed sensing in the signal processing community. Computing the optimal solution involves searching over all the $\binom{M}{K}$ possible combinations and thus is NP-hard. Approximation algorithms are often used. Here, we use orthogonal matching pursuit (OMP) [31]



(a) Spatial pyramid with partition 1-2-3. (b) Temporal pyramid with partition 1-2-4. (c) The smallest possible cell in the given spatio-temporal pyramid.

Fig. 2. Schematic illustration of the partition of data in the spatial, temporal and spatio-temporal pyramid. The size of a cell in which features are pooled at each pyramid level is marked with a color. The left image visualizes the 3-level spatial partition of a single frame into C_s cells. This setup is used to compute the HMP features. The matrix is divided into $S = 1 + 2^2 + 3^2$ cells giving the HMP of the size $14 \times M$, where M is the size of the codebook. The middle image presents a partition of the tactile sequence into C_t cells for the 3-level temporal pyramid $1 - 2 - 4$ giving $T = 1 + 2 + 2^2$. Finally, the dimensionality of the ST-HMP feature vector F_P obtained for the spatio-temporal pyramid pooling is equal to $14 \times 7 \times M$.

to compute the sparse code x. As a greedy-style algorithm, OMP selects the codeword best correlated with the current residual at each iteration, which is the reconstruction error remaining after the codewords chosen thus far. At the first iteration, this residual is exactly the observation y. Once the new codeword is selected, the observation is orthogonally projected onto the span of all the previously selected codewords and the residual is recomputed. The procedure is repeated until the desired K codewords are selected.

Then, the codewords of the codebook are updated one at a time, resulting in a new codebook. This new codebook is then used in the next iteration to recompute the sparse code matrix followed by another round of codebook update. To avoid introducing new non-zero entries in the sparse code matrix X, the update process only considers observations that use the m-th codeword. It can be shown that each iteration of codebook learning followed by updating decreases the reconstruction error given in Eq. 1. In practice, K-SVD converges to good codebooks for a wide range of initializations.

B. Spatio-Temporal Pyramid

Given a tactile sequence, each pixel is represented by the sparse codes computed from a small neighborhood around it, e.g. a patch 4×4 pixels. Spatio-temporal pyramid max pooling is then applied to these sparse codes to generate feature vectors. It partitions the tactile sequence P into spatio-temporal cells C_{st}. The features of each spatio-temporal cell are the max pooled sparse codes, which are the component-wise maxima over all sparse codes within a cell:

$$F(C_{st}) = \left[\max_{j \in C_{st}} |x_{j1}|, .., \max_{j \in C_{st}} |x_{jm}|, .., \max_{j \in C_{st}} |x_{jM}| \right]$$

Here, j ranges over all entries in the cell, and x_{jm} is the m-th component of the sparse code vector x_j of entry j. Note that $F(C_{st})$ has the same dimensionality as the original sparse codes but may be less sparse due to the max pooling operation. The feature F_P describing the whole tactile sequence is the concatenation of aggregated sparse codes in each spatio-temporal cell

$$F_P = \left[F(C^P_{11}), \cdots, F(C^P_{st}), \cdots, F(C^P_{SxT}) \right]$$

Fig. 4. (*Top*) Five robot hands used to collect databases described in Section IV-A. (1) Model of the 3-finger Schunk Dexterous Hand (SDH), (2) 3-finger SDH, (3) 3-finger SDH with Weiss tactile sensors, (4) 2-finger Schunk Parallel Gripper, (5) 5-finger iCub hand. (*Bottom*) Example of grasp execution with the SDH applied to the *Can Cylinder* (SD-5 database). Figures are reproduced with permission from [8] [23] [29].

where $C^P_{st} \subseteq P$ is a spatio-temporal cell generated by spatio-temporal pyramid partitions over the patches, S is the total number of spatial cells and T is the total number of temporal cells. Thus, dimensionality of the feature vector F_P is equal to $S \times T \times M$, where M is the size of the codebook. An example of creating the pyramid is given in Fig. 2.

The main idea behind spatio-temporal pyramid pooling is to achieve different levels of invariance to local deformations, thereby increasing the discrimination of the features. We additionally normalize the feature vectors F_P by their L_2 norm $\sqrt{\|F_P\|^2 + \epsilon}$, where ϵ is a small positive number to make sure the denominator is larger than zero.

Finally, we underline that a dictionary can be learned not only for 2D spatial patches sampled from individual frames, but also for 3D patches (cubes) sampled from blocks of consecutive frames creating a *spatio-temporal dictionary*. We experimentally evaluate this solution in Section IV-C.4.

IV. Experimental Evaluation

We analyzed the performance and properties of the Spatio-Temporal HMP for several synthetic and real databases with diverse characteristics. We considered two typical classification problems for tactile data: grasp stability assessment and

(a) SDS (b) SD-5 (c) SD-10, SPr-10 (d) SPr-7 (e) iCub-10

Fig. 3. Objects used to collect six tactile databases described in Section IV-A. (a) Object and grasp type classes for the Schunk Dextrous Synthetic database; (b) Schunk Dextrous database with 5 real objects: *White Bottle, Black Cylinder, Bleach Cylinder, Spray Bottle, Can Cylinder*. Two types of grasp were applied to the objects (side and top) creating 8 object-grasp pairs; (c) Schunk Dextrous and Schunk Parallel databases with 10 objects: *rubber ball, balsam bottle, rubber duck, empty bottle, full bottle, bad orange, fresh orange, joggling ball, tape, wood block*; (d) Schunk Parallel database with 7 deformable objects: *grape, kiwi, lime, mushroom, orange, plum and tomato*; (e) iCub database with 10 objects: *3 bottles, 2 soda cans, teddy-bear, monkey, lotion and book*. Figure are reproduced with permission from [8] [23] [29]

Database	Data Type	Hand and Sensors			#Obj.	#Grasp Types	Seq. Length		Previous work	
		Type	#Fingers	Size			Avr (RSD)	f [Hz]	Ref.	Representation
SDS	synthetic	Schunk Dextrous Model	3	$12x6$	3	3	156 (74%)	–	[21] [32]	*F*: 1st & 2nd order moment; size & center of contact area *CL*: HMMs, AdaBoost, SVM (rbf kernel)
SD-5	real	Schunk Dextrous	3	$13x6$, $14x6$	5	2	274 (66%)	–	[21] [8]	
SD-10	real	Schunk Dextrous	3	$13x6$	10	-	347 (7.4%)	100	[23]	*F*: 1st & 2nd order moment *CL*: Dynamic Time Warping
SPr-10	real	Schunk Parallel	2	$8x8$	10	-	512 (1.5%)	100	[23]	
SPr-7	real	Schunk Parallel	2	$8x8$	7	-	406 (0.9%)	100	[23]	
iCub-10	real	iCub Hand	5	$1x12$	10	-	12 (29%)	10	[29]	*F*: 1st-3rd moment; min&max press. *CL*: C4.5, SVM(rbf), FS+SVM(rbf), GP+STORK kernel

Tab. 2. Summary of the properties of six databases described in Section IV-A. Abbreviations: *F* - Features, *CL* - Classifier, *GP-STORK* - Gaussian Process with Spatio-Temporal Online Recursive Kernel, *SVM-FS* - Support Vector Machines with features selected using a genetic algorithm, *C4.5* - Decision tree.

Id.	Object & Grasp	HMP		MV-HMP		ST-HMP		HMM_{ERG}	HMM_{LR}	SVM+AB
		Av Acc	σ	Av Acc	σ	Av Acc	σ	Av Acc	Av Acc	Av Acc
1	Black Cylinder Side	71.4%	3.4%	75.7%	15.1%	**99.3%**	2.6%	98.0%	99.0%	-
2	White Bottle Side	67.7%	2.7%	86.0%	8.4%	**99.0%**	3.2%	**99.0%**	98.0%	-
3	White Bottle Top	78.4%	4.7%	87.5%	17.7%	**100%**	0%	97.0%	96.0%	-
4	Black Cylinder Top	69.5%	4.7%	95.0%	8.1%	**100%**	0%	90.0%	93.0%	-
5	Bleach Cylinder Side	82.3%	7.4%	93.3%	11.6%	**100%**	0%	97.0%	98.0%	-
6	Sprinkler Bottle Side	82.7%	4.2%	87.5%	8.1%	**100%**	0%	90.0%	93.0%	-
7	White Bottle	87.5%	7.5%	97.5%	7.9%	**100%**	0%	59.5%	69.0%	73.5%
8	Can Cylinder	83.2%	2.9%	99.3%	2.6%	**100%**	0%	82.0%	86.5%	90.3%
Total Average Accuracy		77.8%		90.2%		**99.8%**		89.1%	91.6%	

Tab. 3. Results for the grasp stability assessment task for the DS-5 database. Comparison of the HMP-based representations with the previously reported results for the ergodic HMM, the left-to-right HMM and the SVM+AdaBoost [21, Table IV] [8, Table II & VII].

Database	Previous work		Discussed in Section IV-C.3						Discussed in Section IV-C.4						
			HMP		MV-HMP		ST-HMP		n_{FD}	HMP_{TD}		MV-HMP_{TD}		ST-HMP_{TD}	
	Av Acc	σ	Av Acc	σ	Av Acc	σ	Av Acc	σ		Av Acc	σ	Av Acc	σ	Av Acc	σ
SD-5	–	–	87.0	3.7	90.5	1.8	**98.9**	2.99	–	–	–	–	–	–	–
SD-10	92.0	9.2	78.7	8.4	**94.0**	7.0	**94.0**	7.0	10	89.2	6.3	**97.0**	6.7	96.0	5.1
SPr-10	89.2	9.1	79.5	10.6	84.5	18.2	88.5	12.9	10	83.0	9.2	85.3	16.6	**91.1**	11.5
SPr-7	91.4	6.9	90.5	6.0	94.3	7.4	**95.7**	6.9	10	92.6	5.0	**97.1**	6.0	**97.1**	6.0

Tab. 4. Comparison of the proposed methods with previous works for an object classification task for four real databases. For the SD-10, SPr-10 and SPr-7 an average accuracy is computed based on confusion matrices for first and second moment [23, Tab. 3, 9, 10].

object instance recognition. Here, we begin with the description of the database and setup used during the experiments.

A. Databases

We evaluated our approach on six different synthetic and real databases on which the state-of-the-art methods have been tested [8] [23] [29]. This allows to examine our method on data with widely varying properties. The databases were collected for objects of diverse shape and physical characteristics using five robot hands as presented in Fig. 3 and 4. We shortly describe each of the databases and summarize the most important information in Tab. 2.

*1) Schunk Dextrous Synthetic (**SDS**) database and Schunk Dextrous database with 5 real objects (**SD-5**) [21] [8]:* These databases were originally collected for the grasp stability assessment problem. Each grasp execution provides a

Database	Previous work								Our work					
	C4.5		SVM^{RBF}		$SVM\text{-}FS^{RBF}$		STORK-TC		n_{FD}	HMP		MV-HMP		ST-HMP	
	Av Acc	σ	Av Acc	σ	Av Acc	σ	Av Acc	σ		Av Acc	σ	Av Acc	σ	Av Acc	σ
iCub-10	98.5	1.2	83.5	6.8	99.5	0.1	99.3	0.2	1	99.4	0.5	99.9	0.5	**100**	0

Tab. 5. Comparison of our methods with previous work for the iCub-10 database [29, Tab. II]. Abbreviations: *C4.5* - Decision tree algorithm, *SVM-FS* - Support Vector Machines with features selected using a genetic algorithm, *STORK-TC*- Gaussian Process with Spatio-Temporal Online Recursive Kernel.

Alignment	HMP		MV HMP		ST HMP	
	Av Acc	σ	Av Acc	σ	Av Acc	σ
Individual	59.7%	3.4%	60.3%	4.0%	71.1%	6.8%
Concatenated	60.9%	3.7%	62.8%	2.9%	**74.4%**	6.8%

Tab. 1. Comparison of different alignments of tactile matrices. Average accuracy over all objects in the SDS database for grasp stability assessment: (a) three $6x12$ pixel matrices (each from a single tactile sensor) are treated individually, and (b) readings from all sensors are concatenated into one $18x12$ pixel matrix.

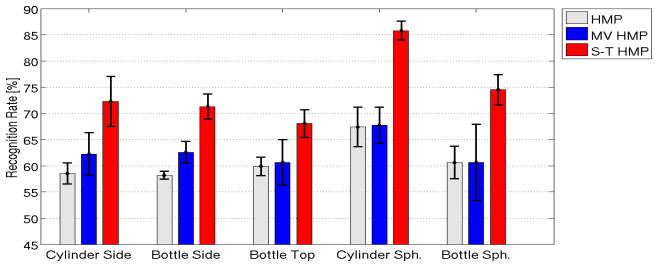


Fig. 5. Results for the grasp stability assessment for the SDS database.

sequence of tactile sensor readings that is labeled as deriving from a stable/unstable grasp. The sequence consists of tactile measurements from the first physical contact of a hand with an object until the fully closed hand configuration is reached or no changes in tactile readings occur for a specified time.

The SDS database was collected for a model of the three finger Schunk Dextrous Hand (SDH) in the RobWorkSim simulator [32]. Each robot finger is equipped with one tactile sensor. The database contains grasps for three objects that are lifted using a side, top or spherical grasp creating five object and grasp type classes (Fig 3(a)). The simulator has a physical model of an object and the robot hand. Synthetic tactile readings are obtained by computing deformation of a sensor surface under force applied to it by the hand. To judge grasp quality, a measure based on the radius of the largest enclosing ball in the grasp wrench space is used.

The SD-5 database was collected for the three finger Schunk Dextrous Hand (SDH) in which each finger is equipped with two tactile sensors. An example of grasp execution for this hand is presented in Figure 4. Figure 3(b) presents five objects from the database to which a top or a side grasp was applied creating eight object–grasp classes listed in Table 3. We used the collection of all publicly available data, namely the class 1-6 from [21] and 7-8 from [8]. To judge grasp stability, the object was lifted and rotated. The grasp was considered successful if the object did not fall or move in the hand. The annotation was done manually. In order to ensure variety of grasps, the initial position of objects was not precisely aligned with respect to the hand. In addition to grasp stability information, the data are annotated with the name of the object to which the grasp was applied. This allows us to also use this database for an object instance recognition task.

*2) Schunk Dextrous and Schunk Parallel databases with 10 objects (**SD-10**, **SPr-10**), and Schunk Parallel database with 7 objects (**SPr-7**) [23]:* These databases were originally collected to compare properties of newly proposed flexible piezoresistive rubber tactile sensors with the popular Weiss sensors [23]. The first two databases were collected for the same set of 10 household objects of complex shape presented in Fig. 3(c). In the case of the SD-10, objects were lifted using a similar hand type as the one used to obtain SD-5, namely the three finger Schunk Dextrous gripper with two tactile sensors per finger (Fig. 4(c)). The SPr-10 data were recorded using the two fingered Schunk Parallel hand with one flexible tactile sensor per finger (Fig. 4(d)). The same robot hand was used to collect the SPr-7 database that contains 7 deformable objects presented in Fig. 3(d). For all three databases grasp execution was similar. An object was manually placed between the gripper jaws. A palpitation procedure started with the first physical contact of the hand with the object. Then, five small squeeze-release steps were executed (fingers were moved back and forward by 1 mm), and finally the gripper released the object.

*3) iCub database with 10 objects (**iCub-10**) [29]:* The database was collected using the five finger iCub hand in which each finger is wrapped with 12 capacitive pressure sensors (Fig. 4(e)). It contains nine everyday objects presented in Fig. 3(e) and one baseline where the grasp was executed with no object. The grasp action was performed by placing an object between the fingers and closing the hand with low velocity. Once, contact was detected, hand was moved with higher velocity to firmly press the object. Action ends when motion was blocked or its trajectory finished.

For more information about the databases, we refer the reader to the corresponding publications listed in Tab. 2.

B. Experimental Setup

Our goal was to encode both spatial and temporal information in the data. To do this, we first applied the original spatial HMP to each frame in a grasping sequence, and then in the second step, added temporal information to the representation. In a simple case, the latter can be done by collecting classification results (labels) of individual frames and applying Majority Voting to recognize the whole sequence. This way the temporal information is used, however the relation between consecutive frames is weekly captured. We refer to this approach as the *Majority Voting HMP (**MV-HMP**)*. In contrast, the ST-HMP combines spatial features from portions of frames over time in hierarchical manner creating a single feature vector for the whole sequence.

Support Vector Machines (SVMs) with a linear kernel were used as a classier in all cases. For rich features provided by sparse coding, this kernel obtains satisfactory results and there is no need to apply more complex distance measures. For each experiment, we report the average recognition rate and standard deviation (σ). For the HMP we report the accuracy for a single frame classified by the SVM and for methods that capture temporal information (MV-HMP and the ST-HMP) for the whole sequence.

1) Grasp Stability Assessment: As in the previous works [21] [8], our experiments were performed for the scenario in which the system has knowledge about an object shape and grasp type (e.g. from visual input). Thus, we train a separate classifier for each pair object-grasp (e.g. *Cylinder-Side*) and report results for each of such pairs separately.

We evaluated our approach on both synthetic (*SDS*) and real data (*SD-5*) closely following the setups from [21] and [8]. For the *SDS* we used a smaller dataset than in [21], since the complete dataset is not publicly available. The detailed information about setups used for grasp stability assessment is presented in Tab. 6 (rows 1-2).

2) Object Instance Recognition: In these experiments, we aimed to distinguish one object from all the others, thus a single classier was trained for all objects. We evaluated our method on five databases (*SD-5, SD-10, SPr-10, SPr-7* and *iCub-10*) that contain objects of various shape, size, weight and stiffness, grasped using four different types of robot hands. In order to form the dataset for the *SD-5*, we used all the sequences corresponding to stable grasps from the real database. We merged side and top grasps for each object. For example, the class *Black Cylinder* contains all stable grasps from *Black Cylinder Side* and *Black Cylinder Top*. This way we increased variability of the data in each class. The detailed information about experimental setups for all databases used for object instance recognition is presented in Tab. 6 (rows 3-7).

C. Experimental Results

In our experiments, we first determined the structure of the learned features for the grasp stability assessment using synthetic data. Given the structure, we performed a full set of experiments on the same data and compared the performance of the ST-HMP with the original HMP and MV-HMP. Next, we utilized several real databases to quantitatively compare the ST-HMP with the state-of-the-art methods that use different types of pre-defined features and encode temporal information in various ways. Then, to demonstrate universal properties of our method, we applied it without any essential changes to data of various characteristics and to two different tasks, grasp stability assessment and object classification. Finally, we improved performance of our method further using a different method of dictionary learning.

1) Structure of the Learned Features: Usually, several tactile sensors are mounted on a robot hand providing multiple readings in parallel. One approach can be to apply the HMP to each reading separately and then fuse feature vectors before classification. The other is to concatenate the readings at the data level, forming one large matrix for HMP. In the first experiment, we compared the two approaches for HMP, MV-HMP and ST-HMP on the SDS synthetic database. All remaining parameters of the methods were pre-selected on a small validation set. Our experiments showed that concatenating readings is by far superior to analyzing them separately (see Tab. 1). This result can easily be explained by the fact that concatenating the readings allows learned features to reflect the dependencies between multiple sensors.

Next, we compared the behavior of the method for various numbers of spatial layers of the HMP algorithm. In our experiments, a one-layer spatial approach provided satisfactory results for all versions of the HMP descriptors and there was a minor gain in performance when a second layer was added. We reason that due to small resolution and dimensionality of inputs, statistics that well represent the data are already captured at the first layer. All further experiments were performed for the concatenated readings and the one-layer spatial HMP. To obtain data for dictionary learning, we grid sampled the input matrix with spatial patches of $4x4$ pixels.

Then, we analyzed the influence of the number of pyramid levels in the temporal dimension. Results are shown in Fig. 8(b). Using multiple levels allows the algorithm to capture information at different time scales and adapt to processes of different temporal resolution. Our baseline is set by pooling features over the whole sequence (indicated as 1). In such a case the ST-HMP already significantly outperforms the MV-HMP with accuracy 67.2% and 60.6% respectively. When more levels are used and the sequence is divided into sub-parts, the recognition rate increases until a point when the level of detail is too high (e.g. 1-2-4-8-16-32) and recognition rate starts to drop.

Finally, for the determined structure, we performed a detailed comparison of the three methods of representing tactile data in which: (a) only spatial information carried by data is encoded (HMP), (b) temporal information is added by aggregating classification results for individual frames using majority voting (MV-HMP), (c) a single spatio-temporal descriptor is created for a whole sequence (ST-HMP). Figure 5 summarizes recognition rates obtained for the synthetic database. Results are consistent across all object-grasp pairs confirming that including temporal information by spatio-temporal pooling is beneficial. The ST-HMP outperforms other methods by a very large margin. The MV-HMP does not increase accuracy for all objects.

Figure 8(a) presents detailed results for these three methods for different size of the spatial dictionary. We see that the ST-HMP constantly outperforms other methods regardless of that parameter value. A relatively small dictionary already provides an acceptable recognition rate due to low dimensionality of the data (16 for $4x4$ patches).

2) Quantitative Evaluation: Using real data, we performed a quantitative comparison of our approach with the state-of-the-art results obtained using different techniques for encoding temporal information, such as Hidden Markov Models (HMMs) [21] [8], Gaussian Processes (GP) with recursive kernels [29], Dynamic Time Warping (DTW) [23],

decision trees [29], and techniques based on SVMs and AdaBoost [8] [29]. As shown in Tab. 2, in all these cases *pre-defined* features have been used, such as higher order moments or geometrical properties of contact regions.

The ST-HMP provided superior or equal performance as the previously published methods for all the analyzed cases of classification tasks and types of data (multiple objects and robot hands). As presented in Tab. 3, the ST-HMP outperformed the HMMs on average by 8.2% for the SD-5 and grasp stability assessment. Similarly, for object recognition our method yielded better accuracy than the DTW for the SD-10 and SPr-7, 4.3% and 2.0% respectively (see Tab. 4) and was almost equally accurate as DTW for SPr-10 (misclassified sequences contain up to 25% of empty frames). Our method obtained perfect recognition for the iCub-10 reaching accuracy of the GP (STORK-TC) and the SVM with feature selection based on a genetic algorithm, and improved upon a decision tree technique, while solely using the SVM with a simple linear kernel (see Tab 5).

These results validate our approach to analyzing series of data. Moreover, our method built on top of unsupervised learned features outperformed all the methods using the manually designed features. It is important to notice that ST-HMP outperforms also the original HMP descriptor proving that capturing temporal information can greatly increase the accuracy in real world environments.

3) Versatility of the Method: The ST-HMP consistently obtained excellent performance when evaluated on the six different synthetic and real databases of diverse properties (SDS, SD-5, SD-10, SPr-10, SPr-7 and iCub-10). These databases were collected for multiple robot hands (see Fig. 4) and a number of objects that manifest a wide range of physical characteristics (shape, size, weight and stiffness, see Fig. 3). Moreover, our method achieved high precision for two different applications without a change in the design.

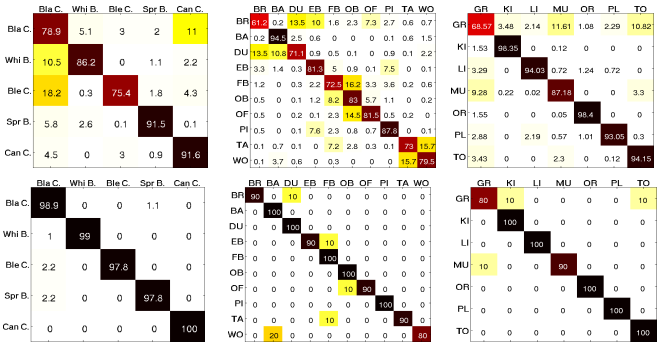
Figure 6 presents the confusion matrices for the object recognition task. The HMP misclassified object instances of similar shape, such as *Black Cylinder* and *Can Cylinder* (SD-5), material *Rubber Duck* and *Rubber Ball* (SD-10) or stiffness *Grape* and *Mushroom* (Pr-7). Adding temporal information about the process of hand alignment around an object allows for almost perfect discrimination between these physically similar objects.

4) Spatio-temporal Dictionary: A dictionary can be estimated not only for 2D patches extracted from individual frames, but also for 3D patches (cubes) sampled from blocks of consecutive frames creating a *spatio-temporal dictionary (**ST-dic**)*. We tested that approach with cubes of a size of $4 \times 4 \times n_{FD}$ where $4 \times 4 \times 1$ indicates *spatial* 2D dictionary used in the previous sections.

Figure 7 shows the relative change in recognition rate for the HMP and ST-HMP between the case of using a spatial 2D dictionary ($n_{FD} = 1$) and ST-dic with increasing amount of available temporal information ($n_{FD} = \{5, 10, 15, 20\}$). For real data and an object recognition task, adding temporal dimension gradually improved accuracy for the HMP. However, experiments with the grasp stability and synthetic

Database	Task	#Cl	#Clf	# Sequances		Data Split	#It
				Class	Total	Train÷Test	
SDS	GS	2	5	150	300	80÷20%	10
SD-5	GS	2	8	25-50	50-100	90÷10%	10
SD-5	OC	5	1	25-50	360	90÷10%	10
SD-10	OC	10	1	10	100	90÷10%	10
SPr-10	OC	10	1	6-13	97	90÷10%	10
SPr-7	OC	7	1	10	70	90÷10%	10
iCub-10	OC	10	1	20	200	80÷20%	30

Tab. 6. Setups for grasp stability assesment (GS) and object classification (OC) experiments presented in Section IV-C. Abbreviations: *#Cl* - number of classes per experiment, *#Clf* - number of classifiers (e.g. one for each object-grasp pair), *#It* - number of times an experiment was repeated.



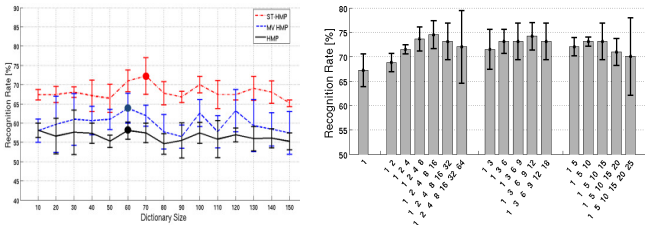
(a) SD-5 (b) SD-10 (c) SPr-7

Fig. 6. Confusion matrices for three databases SD-5, SD-10 and SPr-7 for object recognition. Abbreviations for object classes are given in Fig. 3.

data showed that in rare cases some values of n_{FD} may slightly lower performance of the HMP ($< 1\%$). In contrast, for all analyzed cases the recognition rate for the ST-HMP was improved and increases monotonically with n_{FD} until a point when information captured by the ST-dic and spatio-temporal pyramid become redundant. A summary of these results is given in Tab. 4.

V. Conclusions and Future Work

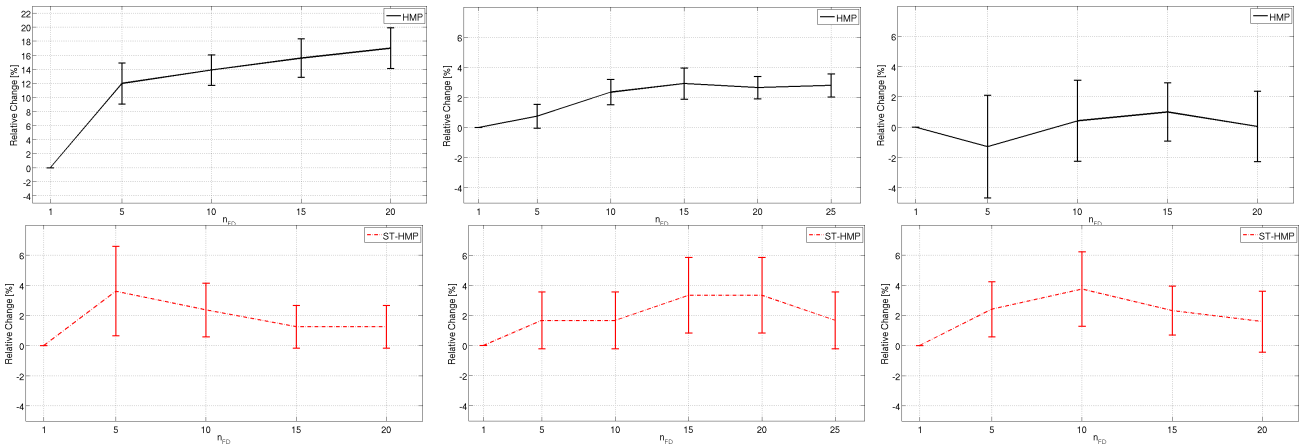
We presented the Spatio-Temporal HMP (ST-HMP), an unsupervised feature learning approach for extracting discriminative structures from sequences of tactile sensor data. Our experiments demonstrated that the ST-HMP outperforms by a large margin approaches ignoring the temporal component as well as previous works using hidden Markov models, Gaussian processes and dynamic programming to



(a) Recognition rate for different sizes of the spatial dictionary for the *Cylinder Side* class. Dots mark the best results that are summarized in Fig. 5.

(b) Recognition rate for different structures of the temporal pyramid. Numbers represents the number of parts into which a sequence is divided at consecutive pyramid levels.

Fig. 8. Analysis of different parameters of the proposed methods.



(a) SD-10 (b) SPr-7 (c) SDS

Fig. 7. Relative change in recognition rate and 95% confidance value for the HMP (top) and ST-HMP (bottom) between the case of using a spatial dictionary ($n_{FD} = 1$) and a spatio-temporal dictionary ($n_{FD} = 5, 10, 15, 20$ frames).

integrate data over time. Our approach leads to state-of-the-art performance for grasping stability assessment and object instance recognition. An extensive evaluation on several synthetic and real databases showed that the ST-HMP is a universal method that can be successfully applied to tactile data originating from different robot hands and objects.

The ability of the ST-HMP to learn rich feature representations from raw temporal data streams makes it a very promising approach for further research in sensor-based object grasping and manipulation. For instance, in addition to tactile data, there exist other useful modalities such as finger positions and joint angles for robot grasping. The ST-HMP has the potential to handle such modalities in a unified framework. In the future, we plan to investigate a principled way to combine multiple modalities and make the ST-HMP applicable to such multi-modal data.

VI. Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research, the EU FP7 project RoboHow.Cog (FP7-ICT-288533) and the Intel Science and Technology Center for Pervasive Computing (ISTC-PC). We thank Yasemin Bekiroglu, Alin Drimus, Jimmy Jørgensen and Harold Soh for providing access to the databases.

References

[1] A. Pronobis and P. Jensfelt. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *ICRA*, 2012.
[2] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3D scenes. In *ICRA*, 2012.
[3] A. Aydemir, A. Pronobis, M. Göbelbecker, and P. Jensfelt. Active visual object search in unknown environments using uncertain semantics. *Transactions on Robotics*, 29(4), 2013.
[4] M. Bjorkman, N. Bergstrom, and D. Kragic. Detecting, segmenting, tracking unknown objects using multi-label MRF inference. *CVIU'14*.
[5] M. Madry, C. H. Ek, R. Detry, Kaiyu Hang, and D. Kragic. Improving generalization for 3D object categorization with Global Structure Histograms. In *IROS*, 2012.
[6] S. Chitta, M. Piccoli, and J. Sturm. Tactile object class and internal state recognition for mobile manipulation. In *ICRA*, 2010.
[7] M. Okamura and M. R Cutkosky. Haptic exploration of fine surface features. In *ICRA*, 1999.
[8] Y. Bekiroglu, J. Laaksonen, J. Jorgensen, V. Kyrki, and D. Kragic. Assessing grasp stability based on learning and haptic data. *TROB'11*.
[9] Hao Dang and P. K. Allen. Stable grasping under pose uncertainty using tactile feedback. *AUTON ROBOT*, 2013.
[10] K. Hsiao, S. Chitta, M. Ciocarlie, and E.G. Jones. Contact-reactive grasping of objects with partial shape information. In *IROS*, 2010.
[11] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard. Object identification with tactile sensors using bag-of-features. In *IROS*, 2009.
[12] N. Gorges, S. E. Navarro, D. Goger, and H. Worn. Haptic object recognition using passive joints and haptic key features. In *ICRA'10*.
[13] D. Xu, G. E. Loeb, and J. A. Fishel. Tactile identification of objects using Bayesian exploration. In *ICRA*, 2013.
[14] M Madry, D. Song Song, and D. Kragic. From object categories to grasp transfer using probabilistic reasoning. In *ICRA*, 2012.
[15] Y. Bekiroglu, D. Song, L. Wang, and D. Kragic. A probabilistic framework for task-oriented grasp stability assessment. In *ICRA*, 2013.
[16] Kai Yu, Yuanqing Lin, and J. Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011.
[17] L. Bo, X. Ren, and D. Fox. Multipath sparse coding using hierarchical matching pursuit. In *CVPR*, 2013.
[18] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *NC*, 2006.
[19] M. Schopfer, M. Pardowitz, R. Haschke, and H. Ritter. Identifying relevant tactile features for object identification. *STAR*, 2012.
[20] L. Bo, X. Ren, and D. Fox. Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms. In *NIPS*, 2011.
[21] Y. Bekiroglu, D. Kragic, and V. Kyrki. Learning grasp stability based on tactile data and hmms. In *RO-MAN*, 2010.
[22] R. A. Russell. Object recognition by a smart tactile sensor. In *ACRA*, 2000.
[23] A. Drimus, G. Kootstra, A. Bilberg, and D. Kragic. Design of a flexible tactile sensor for classification of rigid and deformable objects. *RAS*, 2012.
[24] G. Heidemann and M. Schopfer. Dynamic tactile sensing for object identification. In *ICRA*, 2004.
[25] Z. Pezzementi, E. Plaku, C. Reyda, and G.D. Hager. Tactile-object recognition from appearance information. *TOR*, 2011.
[26] Hongbin Liu, J. Greco, Xiaojing Song, J. Bimbo, L. Seneviratne, and K. Althoefer. Tactile image based contact shape recognition using neural network. In *MFI*, 2012.
[27] Z. Pezzementi, C. Reyda, and G. D. Hager. Object mapping, recognition, and localization from tactile geometry. In *ICRA*, 2011.
[28] Hongbin Liu, Xiaojing Song, T. Nanayakkara, L. D. Seneviratne, and K. Althoefer. A computationally fast algorithm for local contact shape and pose classification using a tactile array sensor. In *ICRA*, 2012.
[29] H. Soh, Y. Su, and Y. Demiris. Online spatio-temporal gp experts with application to tactile classification. In *IROS*, 2012.
[30] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Algorithm for designing overcomplete dict. for sparse representation. *SP*, 2006.
[31] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *SSC*, 1993.
[32] J. Jorgensen, L. Ellekilde, and H. Petersen. RobWorkSim - an open simulator for sensor based grasping. In *ISR*, 2010.