Modeling Target-Distractor Discrimination for Haptic Search in a 3D Environment

Alexandra Moringen*, Witali Aswolinkiy*, Gereon Büscher*, Guillaume Walck*, Robert Haschke*, and Helge Ritter*

***Abstract*— The ability to discriminate between target and distractors, using the information perceived by the hand over time, is essential to perform haptic search successfully, be it for a human hand or a suitably sensorized anthropomorphic robot hand. To address the latter, we train a binary classifier to perform this discrimination during unconstrained haptic search performed by sighted study participants who were blindfolded. In this work, we test different representational concepts and compare the results with the human classification performance. This approach both guides our understanding of human haptic interaction with the 3D environment and aids future modeling of artificial touch for anthropomorphic robot hands. Our contribution is three-fold. Firstly, we are able to acquire a synchronized multimodal time series of exceptionally high spatio-temporal resolution of both the 3D environment and the hand with our novel experimental setup. It includes our Modular Haptic Stimulus Board to represent a 3D environment and a novel tactile glove equipped with position tracking markers and joint angle sensors. Secondly, we introduce a machine learning approach inspired by a novel application of the feature guidance concept for vision (Wolfe et al., 2007 [1]) to modeling of haptic search in a 3D environment, focusing on the target-distractor discrimination. Finally, we compare results for two different types of artificial neural networks, a feed-forward and a recurrent network. We show that using recurrent networks, and therefore integrating information over time, improves the classification results. The evaluation also shows that classification accuracy is the highest for the combination of both the tactile and the joint angle modalities.**

I. INTRODUCTION

Searching for something is a very common task for humans. Vision-based search is an efficient solution humans exploit most of the time, but haptic search in a 3D environment is another skill that they master to find object in an efficient manner. One example is finding keys among objects in a pocket just by using the sense of touch, while the vision is focused on solving navigation to approach the door.

While the complex task of vision-based search has been implemented in robotic grasping by further improving segmentation, classification and recognition algorithms, this approach can not be employed in all situations. Occlusion or invisible parts of objects require action such as change of view point, either by moving the camera, removing occluding objects or by rotating the target object. Haptic exploration could immediately help find, classify or even refine the model of the invisible part of the target object.

The technological development in the past years has gradually enabled us to also acquire both the tactile and kinematic data for a multi-fingered robot hand, and explore the extremely complex but efficient haptic search possibilities. The newest development include tactile flesh and tactile finger nails [2], [3], which, even with low resolution, open the way to interesting haptic exploration with robot hands. Moreover, the existing data acquisition systems capturing human hand motion and tactile interaction currently offer similar resolution and coverage as the sensors in robotic systems[1].

Therefore, a human-inspired model of haptic search employed on an anthropomorphic multi-fingered robotic platform becomes feasible. It is advantageous for both robots' own interaction with its environment, and a better understanding and prediction of human actions by a robot in a HRI-scenario. Running a model on a robot or in simulation is then the best proof of concept of the haptic search model efficiency. However, both modeling of a human haptic search strategy in a three-dimensional environment and its application on robots have yet not been tackled due to a number of challenging issues involved.

In order to perform haptic search we need to model the following two fundamental functionalities, tightly coupled with each other: an algorithm to guide the hand through the environment during the target candidate selection process, and a classifier that, based on the acquired data, is able to differentiate between a target of search and a distractor. The non-hierarchical interleaved interaction of both functionalities with each other strongly resembles the *strange loop* discussed by Hofstadter [4], and leads us to a wide range of fascinating questions, from which we will focus on two major ones:

What is the optimal representation of a target and a distractor object that enables the most efficient guidance of the search process? What information is essential for a decision to move to a new search location? In particular, how should a switch between exploration of different distractor objects be accommodated for in a model that integrates over time? These questions have not been addressed in the literature for a three-dimensional search scenario. Note that the separation into two components, the target-distractor classification and the search strategy is solely used for description purposes.

From the above questions follows the necessity to find a suitable platform-independent representation of the data acquired through haptic interaction with the 3D environment.

*Cognitive Interaction Technology (CITEC), Bielefeld University, Bielefeld, Germany, corresponding author A.Moringen abarch@techfak.uni-bielefeld.de

[1]Video of a tactile glove and its comparison with robot hand tactile sensors is available under the following link: https://www.youtube.com/watch?v=LKwOpRUCs7s

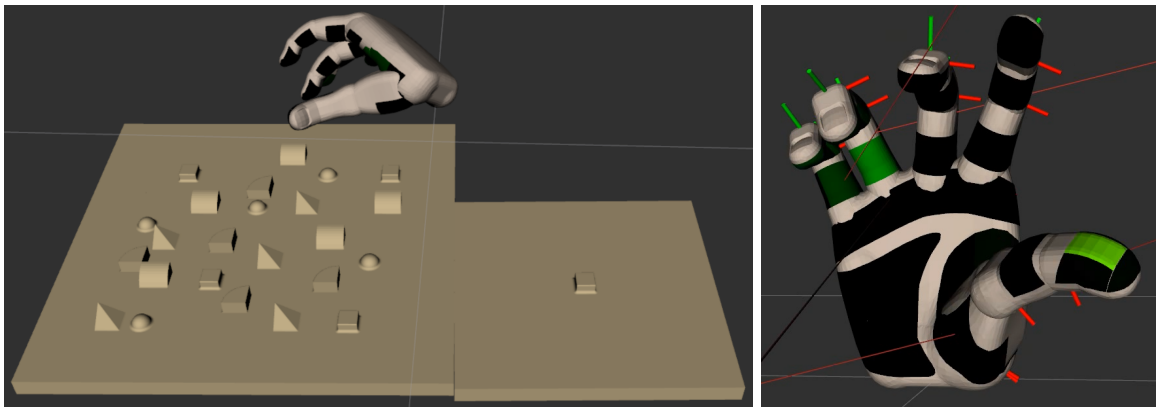


Fig. 1. Snapshots of a full data visualization for the acquired data in an exemplary search scenario (cf. Fig. 2), including the following modalities: position and orientation of the 3D environment Modular Haptic Stimulus Board (MHSB), pose and posture of the hand, and pressure measured by the tactile glove. Left: MHSB and the hand model; Right: Hand model with visualized tactile sensor measurement (in green).

A desirable representation should fulfill the requirements posed by the central goal: to solve the haptic search task in an efficient and robust manner. This fundamental research question is open and has to be solved in order to provide a platform for autonomous robots that learn their own haptic search with e.g. reinforcement learning, or for applications in remote robotics.

In the long run we aim for an autonomous haptic exploration and search performed by a dexterous robot hand fitted with tactile sensors. In this paper, we contribute by evaluating several representational concepts and by verifying their quality in a target/distractor discrimination task based on an exemplary data acquired from human participants. Our approach to tackle the issue of efficiency is inspired by visual search as described in Section II. The 3D world model and the setup employed for the capture of the haptic manual interaction are described in Sections III-B and III-D. To achieve robustness, we investigate a method of multimodal and temporal integration. To this end, we will compare two approaches, one that performs target-distractor classification based only on one point in time with a feed-forward neural network Extreme Learning Machine (ELM, [18]), and one approach that performs temporal integration of the time series with a recurrent Echo State Network (ESN, [19]). These networks are fast to train and allow therefore rapid experimentation. Due to the space constraints, we will only briefly introduce the theory of both ELM and ESN in Section III-F. We restrict the implementation to a case in which the target object class is known. The resulting target/distractor classifier is indispensable for an initial reward calculation during reinforcement learning of a haptic search policy that will be performed in simulation in future work. The results are discussed in Section IV.

II. RELATED WORK

While visual search has been widely investigated and implemented (e.g. [5], [6], [7]), haptic search modeling remains very sparse. In [8] the author describes a simulation of haptic search for patches of different roughness. His work focuses on an application of the guided search theory (GS4) [1] to haptic search w.r.t. the bottom-up and top-down guidance. Martins et al. [9] present a Bayesian model for integration of attentional mechanisms for robotic haptic exploration of surfaces. This work goes in a similar direction and implements haptic exploration as a combination of a stimulus-driven process and a goal-directed modulation. Morash [10] focuses on the *detection radius* as a factor that modulates the strategy employed during haptic search. The conceptual foundation of the present work builds upon a similar idea of the global modulation on the search strategy. However, we build on the hypothesis that the haptic search strategy is modulated by the features of the search target, which is the first application of Guided Search Theory [1] to modeling of haptic search in a 3D environment. Therefore, it remains to be investigated how the modulation by the detection radius and by the target features are exactly related to each other.

We have previously performed a quantitative investigation of the influence of the target object feature for a complex three-dimensional environment [11]. In this work Krieger et al. investigated the intuitive idea that different shape features of the target object such as height, size or curvature, modulate the haptic search strategy. A simple example: Search for a high target object among relatively low objects is most effective at a particular height, determined by the height of the target object. Therefore, building upon the previous work, we base our modeling approach on the following central hypothesis: efficient haptic search, similar to efficient visual search [1], is guided by a small set of guiding features defined by the target object. Haptic guiding features are specific to the haptic modality, such as object height or curvature. We then introduce the novel term *haptic guiding features*, which refers to the features that have an effect on both the top-down as well as the bottom-up guidance during search, as well as the target-distractor discrimination. Following this concept, we use only trials recorded during search for a given target object to train a classifier to discriminate between this target object and the corresponding

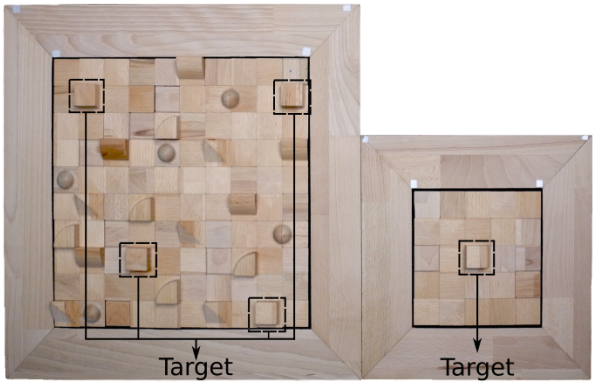


Fig. 2. An example of an experimental setting MHSB used in a trial. The right board presents the target of search (dashed lines) that needs to be memorized in the first step of the task by blindfolded study participants. The left board contains multiple targets distributed among distractors which form the search environment. Distractors are represented by altogether four different object classes. Retroreflective foil (top corners of both boards) is used for estimation of the absolute board position. The figure is taken from [12].

distractors (see Section III-F for more details).

Due to the high complexity associated with a synergy of contact mechanics of the hand [13] and proprioception that takes place during haptic search, an encompassing data acquisition from the hand poses the first challenge for haptic search modeling. The second challenge that needs to be addressed is finding an appropriate modeling approach to a representation of the real world complexity in an experimental setting. In our work we address both challenges with an approach whose great advantage is to enable both acquisition of haptic search data in a scenario of real-world complexity and a straight-forward data transfer to visualization/simulation, exemplified in Figures 2 and 1, respectively.

Most experiments tackling haptic interaction in a 3D environment either deal with a quantitative analysis based on a strongly restricted set of simple objects, search scenario or the degrees of freedom that the study participants are allowed to use, and evaluate the search time for different experimental conditions (e.g. [14], [15]). Another research direction is a qualitative analysis based on a larger set of known objects. In our work we pursue to build a bridge between these two directions and, therefore, employ a framework that enables a modular approach to a construction of a complex three-dimensional environment, the Modular Haptic Stimulus Board[2]. An exemplary three-dimensional environment created with this stimulus material and employed in this work is presented in Figure 2. It illustrates the target of search (outlined in the right board) and the search environment in which multiple targets are placed among distractors (left board). Both boards are employed in our experimental setting. MHSB has been used in previous experiments to enable observation and quantitative modeling of haptic interaction [16], [11], [17] (see [16] for a detailed description of the stimulus material). In order to tackle the second challenge, the data acquisition of the manual interaction, we perform a synchronized capture of the most promising modalities – the spatio-temporal pressure profile, the kinematic hand configuration, including joint angles and absolute three-dimensional trajectory of the palm. Figure 1 shows snapshots of the data visualization[3].

[2]Video of the MHSB and its applications is available under the following link: https://www.youtube.com/watch?v=CftpCCrIAuw

Fig. 3. Five object classes that are employed in the experimental setup in both roles in turns, as a target or as a distractor.

III. METHODS

A. Participants

Ten right-handed sighted individuals aged 20-28 participated in the study. Data of four female and six male participants has been postprocessed and has been employed for the evaluation in this paper. The protocol was approved by the Bielefeld University Ethics Committee, and an informed consent was obtained from all participants prior to their participation. None of the participants had any prior knowledge of the experimental design or the stimuli.

B. Stimulus Material and Experimental Scenario

The stimulus material of the MHSB has been previously employed in a range of studies [16], [11], [17], and represents a three-dimensional shape environment through a combination of wooden bricks. Through this design MHSB is striving for a good balance between the ecological validity and the controllability of factors.

In this study, five identical object copies per object class $l \in \{1,\ldots,5\}$ have been used (see Figure 3), altogether 25 bricks with shapes carved on top. In turns, one of the object classes l has been employed as a target and the other object classes $\{1,\ldots,5\} \setminus \{l\}$ as distractors. On the right 5×5-brick board (see Figure 2), only one object serving in the role of the target has been presented to the study participants. The rest of the board has been filled with planar-surfaced neutral bricks. The left 10×10-brick board illustrated in the same figure represented the search environment with all remaining 24 bricks of all five object classes, presenting four occurrences of the target object randomly distributed among four distractor object classes and the neutral bricks.

C. Task and Procedure

The study participants were blindfolded and asked to perform the following three-staged task in the experimental setting previously discussed in Section III-B.

[3]Visualization of a full trial can be found under the following link: https://www.techfak.uni-bielefeld.de/persons/abarch/videos/td_viz.ogv.

1) **Memorize the target object** presented in the small MHSB on the right.
2) **Search** for multiple instances of the target object in the large MHSB on the left (the search environment). Perform the task as fast as possible, and memorize as many positions of the target objects as possible, until the time limit is reached.
3) **Verify the success of the performed search** by going back to the search environment and retrieving the target object placement from memory.

To encourage the study participants to use the most efficient search strategies, the time for the completion of Stage 2 was limited to 30 seconds, but no instructions were given how to approach the objects, squeeze or touch them. Importantly, in Stage 2 it was not allowed to show at the detected target object or say that the object has been found, but Stage 3 permits to recover this information. This restriction permits to avoid leaving any specific artifacts of finding the target object in the recording that could later affect the representations learned by the classifier.

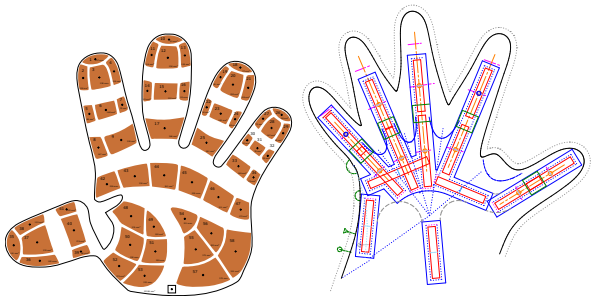
The completion of Stage 3 was limited to 10 seconds to discourage a renewed haptic search. The participants were asked to retrieve the position of the target object from memory serving as a verification of the successful performance in Stage 2. Once a target object candidate was retrieved, the participants were asked to triple tap on the corresponding brick.

An individual recording session took approximately one hour. Each participant performed five trials, whereby in each trial both the target object as well as the distribution of distractors in the search environment has been changed. Before the data recording started, two rehearsal trials have been performed.

D. Experimental Setup and Software Tools

The multimodal time-series representing the dynamics of the hand during haptic interaction as well as the position of the stimuli were synchronously recorded with multiple devices available in the lab [20], including Vicon tracking, tactile RGB glove and camera recording. The recording devices illustrated in Figure 5 will be described in detail in the paragraphs below. On the software side we used hand tracking from point clouds [21] based on an articulated hand model, and an automatic labeling tool [12]. Both employ the three-dimensional Vicon Marker trajectories of the hand and the recorded position of the stimulus material.

1) Glove: The touch sensitive glove used in this study has 58 cells covering a sensitive area of 52% of the palm. Gaps above the joints are retained to increase mobility and to avoid improper values due to self-collision when bending the fingers. The configuration of sensors is shown in Fig.4a). The sensitive cells are composed of several layers of conductive textiles that enable the incoming forces to be read as electrical resistance values. The choice of material provides the sensors with an elastic character. This facilitates tactile properties such as pointy or dull to be felt through the glove. Each cell is scanned at 156 Hz with 12-bit resolution. A detailed evaluation of the predecessor prototype is published in [22].



(a) Tactile sensor distribution (b) Flexion sensor arrangement

Fig. 4. Cut pattern of the multi-modal sensing glove. (a) Palm side with sensitive areas marked in orange. Interspaces are made from elastic breathable mesh fabric. (b) Back of the hand with pockets to fit flexion-sensors of a Cyberglove I.

During the conducted study, the hand posture is tracked and recorded in parallel with two types of sensors. On the one side, palm and finger tracking is performed via the Vicon system (described in the following), which provides absolute position data. On the other side the tracking is performed via 18 flexion sensors of an Immersion Cyberglove I, which provide joint angles at 107 Hz with 8-bit resolution. Occlusion or intersections of markers occasionally causes information loss on Vicon, especially when recording the intricate human hand. The Cyberglove provides data independent of its visibility, but with inferior accuracy.

Wearing the Cyberglove above the tactile glove leads to pre-pressure on the sensors and severely impaired mobility of the hand, which also applies in the case of wearing the Cyberglove underneath. By disassembling the Cyberglove and integrating the flexion sensors into designated pockets of the tactile glove, we were able to create a lightweight, mobile and breathable multimodal sensing glove. The arrangement of the sensors is retained from that of the original Cyberglove, the cutpattern is displayed in Fig. 4b). The integration offers comparatively less movement restriction and perspiration for the human subjects, which should lead to better measurement results. To sum up, the glove offers an exceptionally high spatio-temporal resolution of both, joint and tactile sensors, and was tested for the first time during the described experiment.

2) Vicon: For capturing the position of the hand and the MHSB, the Vicon system was used [4]. It records motion data with a frequency of 200 Hz, using retroreflective markers that are tracked by infrared cameras. Also included is a Basler camera, generating a top-down view for the experiment. Two cameras additionally generated the side-views.

3) Data specification: The acquired data used for the classification contains two modalities and can be specified as follows:

- Joint angles recorded with the Cyberglove sensors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T\}$, with $\boldsymbol{a}_i \in \mathbb{R}^{18}$
- Tactile measurements $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T\}$, with $\boldsymbol{s}_i \in \mathbb{R}^{58}$.

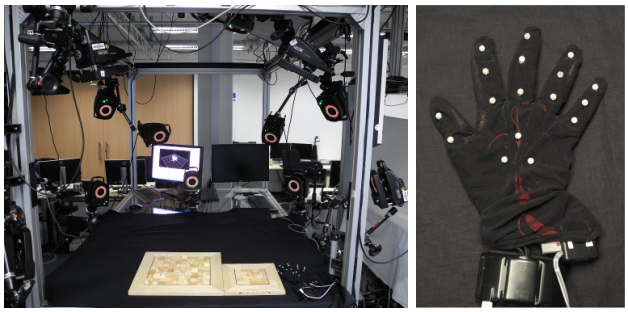


Fig. 5. Recording setup contains a wide range of devices: Vicon cameras, glove with integrated joint and tactile sensors, USB and Basler cameras for verification (left). Placement of the retroreflective Vicon markers on the glove, employed for finger and hand tracking (right).

The above acquired time-series are synchronized based on the recorded time stamps and merged resulting in $\{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_T\}$ with $\boldsymbol{m}_i \in \mathbb{R}^{76}$. In the following, to simplify the notation, we will denote the recorded data by $\boldsymbol{x}$ as a stand-in for the variables defined above.

Due to strong differences between the study participants, the data has been normalized with a participant-wise z-transform prior to the classification training.

4) Auto-labeling and real-time auto-tracking: Because the trajectory data acquired with the help of the Vicon system is not gap-free, we employ an auto-tracker that fits a hand model into the point cloud to fill in the trajectory gaps [21]. Based on the resulting trajectories, the automatic pointwise labeling of the time-series [12] provides ground truth for the object classes that are being explored during the corresponding point in time. As a result, for a multimodal time-series $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\}$ that represents one haptic search trial, the auto-labeling provides the matching labels $\{y_1,\ldots,y_T\}$ corresponding to the sequence of explored objects. There is no intermediate level labeling, extracting which finger, or which motion is applied, only the class of the object is known for the time-series. The procedure that we use is a heuristic-based estimate that receives the object placement on the board and the position of the hand- and finger-markers as an input. This labeling is fully automatized, and no time-costly manual annotation is needed. The partitioning and learning of the resulting labeled time-series data $\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_T,y_T)\}$ will be described in Section III-F.

E. Data visualization with RVIZ

To enhance post-processing and analysis of the acquired data, both the MHSB and the tactile glove data were visualized in RVIZ. Firstly, the configuration of the wooden blocks assembly is described using the *Unified Robot Description Model* (URDF), automatically generated by a combination of a XACRO template file and a YAML configuration file containing the encoding of wooden blocks placement matching the real board. This URDF can not only serve for rendering the look of the board in RVIZ (Fig. 1), but also provide a physical model (collision object) in a simulation environment such as Gazebo for future simulated exploration with a robotic hand. Secondly, the pose, posture and touch data of the tactile glove were displayed thanks to a human hand URDF including markers in the shape of the cut pattern projected on the meshes of the hand 3D model. These markers change color dynamically according to tactile data.

F. Classification Approach

We propose to train target class-specific binary classifiers to discriminate between the target object and the distractor items. The binary classification approach follows from the assumption that during search no identification of individual distractor classes is performed. The target class specific classification approach described in detail below, follows from the hypothesis that the target features modulate the distractor exploration. Therefore, for a given class l of the target object, we will train a corresponding classifier C_l, which will determine, whether a given data point belongs to the target object or a distractor. The data points corresponding to the exploration of all distractor classes will yield the set of negative examples, and the data points corresponding to exploration of the target will yield the set of positive examples. Because each one of the ten considered study participants conducted one trial per target object class, we obtain ten subsets of both types (target subset and distractor subset) for each one of the five target object classes l. To formalize the above, given a target object l, we define X_p^l to be the time-series corresponding to the target object exploration by a study participant p during search in the corresponding trial. By iterating over all study participants, we obtain the cumulative data set that characterizes haptic exploration of l in the role of a target object during the haptic search: $X_{\mathscr{T}}^l := \bigcup_{p=1}^{P} X_p^l$, where $P = 10$ is the total number of study participants. The complementary set corresponding to the distractor exploration with respect to a given target object class l encompasses the time-series segments that correspond to exploration of all object classes apart from l denoted by $\bar{l}$: $X_{\mathscr{D}}^l := \bigcup_{p=1}^{P} X_p^{\bar{l}}$. Importantly, X_p^l and $X_p^{\bar{l}}$ both contain data from the same trial defined by the task to find object l among distractors $\bar{l}$. Note that within both sets we build segments corresponding to the interaction with a particular object l located on the board. Each set $X_{\mathscr{T}}^l \cup X_{\mathscr{D}}^l$ contains approx. 60.000 points.

The quality of the classifier depends heavily on the division of the time series into training and testing subsets. The simplest way is to divide the points from the time series randomly. This point-wise partitioning, however, may unrealistically simplify the learning task, since the classifier may just learn to interpolate between the given points. More challenging is a segment-wise division, where the trial data is divided into segments corresponding to the explored object. Then, during testing, the classifier must be able to differentiate between the target and the distractors on all points of a previously unseen segment. The training and testing splits are repeated using the 5-fold cross-validation scheme.

In order to compare classification results for a feed-foward and a recurrent network we employ ELMs and ESNs, respectively. ELMs and ESNs are neural networks with three

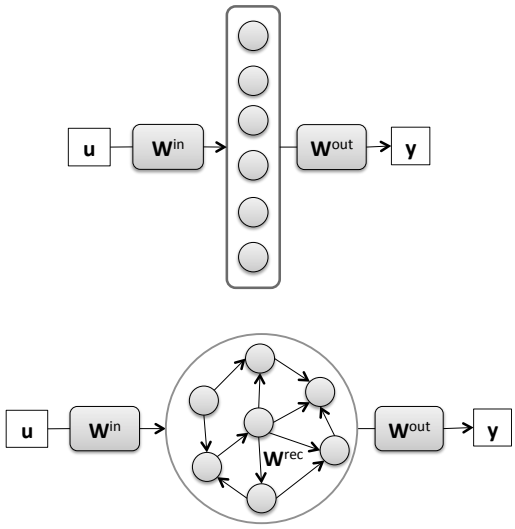


Fig. 6. Architectures of Extreme Learning Machine (top) and Echo State Network (bottom).

layers: An input layer $\boldsymbol{x} \in \mathbb{R}^I$, a hidden layer $\boldsymbol{h} \in \mathbb{R}^N$ with N hidden neurons, and a linear output layer $\boldsymbol{y} \in \mathbb{R}^O$ (cf. Fig. 6). In ELMs, the hidden layer is feed-forward and the output is computed by $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{W}^{out} a(\boldsymbol{W}^{in}\boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{W}^{in} \in \mathbb{R}^{N \times I}$ is the weight matrix from the inputs to the hidden neurons, $a(\cdot)$ the activation function applied element-wise to the neuron inputs, e.g. hyperbolic tangent or logistic, $\boldsymbol{b} \in \mathbb{R}^N$ the neuron biases and $\boldsymbol{W}^{out} \in \mathbb{R}^{O \times N}$ the weight matrix from the hidden neurons to the outputs. $\boldsymbol{W}^{in}$ and $\boldsymbol{b}$ are initialized randomly and remain fixed.

In ESNs, the hidden layer is a reservoir of recurrently connected neurons, which provide a non-linear fading memory of the inputs. The reservoir states $\boldsymbol{h}$ and the readouts $\boldsymbol{y}$ are updated according to $\boldsymbol{h}(k) = (1-\lambda)\boldsymbol{h}(k-1) + \lambda a(\boldsymbol{W}^{rec}\boldsymbol{h}(k-1) + \boldsymbol{W}^{in}\boldsymbol{x}(k) + \boldsymbol{b})$ and $\boldsymbol{y}(k) = \boldsymbol{W}^{out}\boldsymbol{h}(k)$, respectively. $\lambda \in (0,1]$ is the leaking rate and $\boldsymbol{W}^{rec} \in \mathbb{R}^{N \times N}$ is the recurrent weight matrix. $\boldsymbol{W}^{rec}$ is initialized randomly and typically scaled so that the spectral radius of $\boldsymbol{W}^{rec}$ is smaller than one. The size of the reservoir memory depends on the number of the neurons and the leakage rate. Compared to an ELM of the same size, an ESN provides less information about the current time step, but contains information from previous steps.

Training of the networks is restricted to the output layer, which can be trained effectively with ridge regression (closed-form solution). Note that the hidden layer neuron activations for all acquired time series are calculated and stored before the training/testing of the classifiers. An important aspect that needs to be considered during training is the class imbalance: Since the participants spent more time exploring the distractors, there are fewer target data points than distractor points. We balance the classes out during training of the output layer by giving them a higher weight corresponding the ratio of the distractor to the target points: $(\boldsymbol{W}^{out})^T = (\boldsymbol{H}^T\boldsymbol{V}\boldsymbol{H} + \alpha\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{V}\boldsymbol{T}$, where α is the regularization strength, $\boldsymbol{I}$ the identity matrix, $\boldsymbol{V}$ the sample weights containing the weights of the data points in the diagonal and $\boldsymbol{H}$ and $\boldsymbol{T}$ the row-wise collected neuron activations and targets, respectively. For classification, the two classes are encoded with -1 and 1. The winner class is determined by rounding the prediction and taking the class with the closest value. The magnitude of the network output corresponds to the distance to the decision boundary and may be interpreted as the certainty of the classifier in its decision.

To find suitable values for the hyper-parameters (scaling of $\boldsymbol{W}^{in}$ and $\boldsymbol{W}^{rec}$, λ, α, N), a brief parameter search using the procedure termed *heuristic search* and defined in [23] was conducted.

IV. RESULTS

As previously defined, $X^l_{\mathscr{T}}$ and $X^l_{\mathscr{D}}$ denote target and distractor exploration during search of object l, respectively. The training with point-wise randomly shuffled sets $X^l_{\mathscr{T}}$ and $X^l_{\mathscr{D}}$ yields close to 100% classification accuracy in the 5-fold cross validation tests for both the ELM and the ESN (plot is not displayed in this paper). We incline to attribute this to the classifiers simply interpolating between very similar, temporally close-by data. Importantly, in order to test how well our classifier approach is generalizing, we have used randomly shuffled segments generated by the auto-labeling tool, instead of points, to create a train-test data split.

In order to evaluate, *whether temporal integration within the time series is advantageous*, the first evaluation compares classification accuracy between ESN and ELM (see Fig. 7). The x-axis depicts the number of neurons $N = 10, \ldots, 300$, step size equals 10. The y-axis denotes the mean classification accuracy over all five individual target-distractor classifiers based on 5-fold cross validation. The plot illustrates well that the recurrent model ESN performs better than ELM on the whole range. The same advantage of the ESN over the ELM is robustly sustained through the further tests (up to $N = 1500$, not displayed in this paper). Therefore, we can infer that the temporal integration that results from recurrence is advantageous for the classification accuracy. Note that individual ESN target-specific classifiers can be characterized by a large difference in classification accuracy of about 20% between the best and the worst for a given parameter combination. From this we infer that the parameter optimization has to be conducted for each classifier individually. Intuitively, different target objects result in different search strategies that, in turn, require individual structure and parameterization of the corresponding ESNs.

The best accuracy resulting from minimizing the average error over all classifiers we have achieved is c.a. 70%, for an individual classifier the optimization yielded approx. 78%. The result has been calculated for all points including those that are located directly in the beginning of the respective object exploration segment, and presumably do not have sufficient information for a robust classification, even for a human who keeps exploring the same object. A more detailed

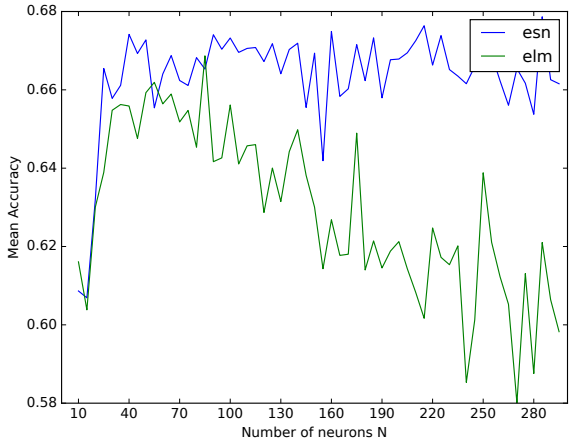


Fig. 7. Comparison between ESN and ELM for an increasing number of neurons N (x-axis). y-axis depicts the mean accuracy estimated over individual target-distractor classifiers.

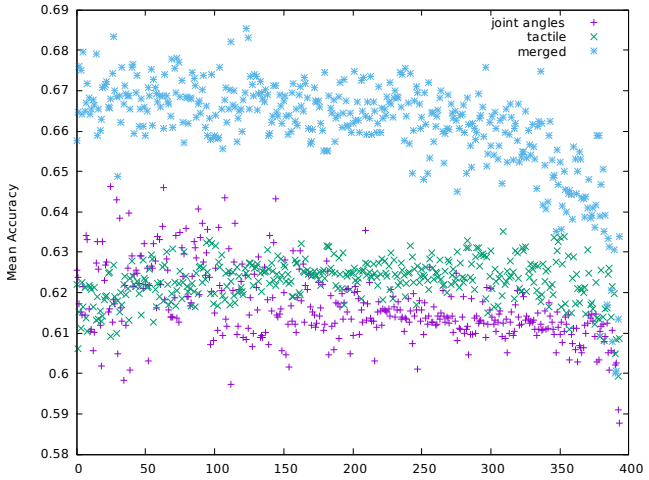


Fig. 8. Comparison between mean accuracy achieved for point-wise classification with different modalities.

analysis of the error dynamics will be performed in the third research question described below.

Our next research question is whether *using multimodal data is advantageous for the classification rate*. Fig. 8 presents a comparison of ESN-classification accuracy for three different types of data: *joint angles*, *pressure profiles* and both modalities together denoted by *merged*. The plot presents the results of 400 evaluations of different values of the ESN parameters, such as leakage rate and the number of neurons. The x-axis denotes the index of a given parameter combination, while the accuracy values (y-axis) are sorted ascending according to the corresponding training error (not displayed in the plot). The plot clearly shows that the combination of both modalities is advantageous for classification in comparison to training the classifier with individual modalities independent of the parameter combination.

In the third evaluation we have pursued to analyze the dynamics of the classification accuracy generated by an ESN

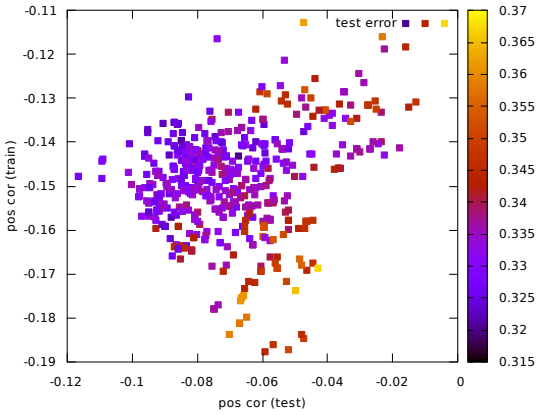


Fig. 9. Biserial correlation on the training and the test sets (x- and y-axis, resp.), the test error illustrated by means of a color palette.

within a segment of the time series. By doing so we address the following question: *Does an ESN make less mistakes as it approaches the segment end and more information about the object being explored become available?* For this purpose, we have evaluated the biserial correlation coefficient r that in general measures the correlation between a binary variable and a continuous variable. Fig. 9 presents correlations calculated between the position of the data point $(1,\ldots,S)$ w.r.t. the corresponding segment border S and the test error for 400 different parameter combinations, $p < 0.05$. It shows that the value of r for both the training and the test data is negative, implying a weak negative relationship between the position within the segment and the occurrence of classification errors. Contrary to the correlation values calculated for the test data (x-axis) that have a positive effect on the test error the lower the value, the correlation value on the training data (y-axis) in the interval $[-0.15,-0.13]$ is associated with the best test results. Our evaluations showed that a stronger correlation on the training data corresponds to an increasing training accuracy which in turn results in overfitting and in decrease of classification accuracy. Following this test we have performed an evaluation of classifier accuracy for the last 15% of the data points within a segment which yielded a large improvement in classification accuracy, which is a highly exciting result. For an individual target-distractor classifier, target class cuboid, Figure 10 illustrates a comparison between the test error (green) and the test error evaluated for the last 15% of the segment points (violet). The best test result for the 15% evaluation reaches the error value of 11%, which is a large improvement in comparison to all presented results. This may imply that towards the end of exploration of an individual object, the human makes better-informed descisions w.r.t. the exploration strategy.

V. Conclusion

Our work directed at the modeling of efficient haptic search in a 3D environment is motivated by our central

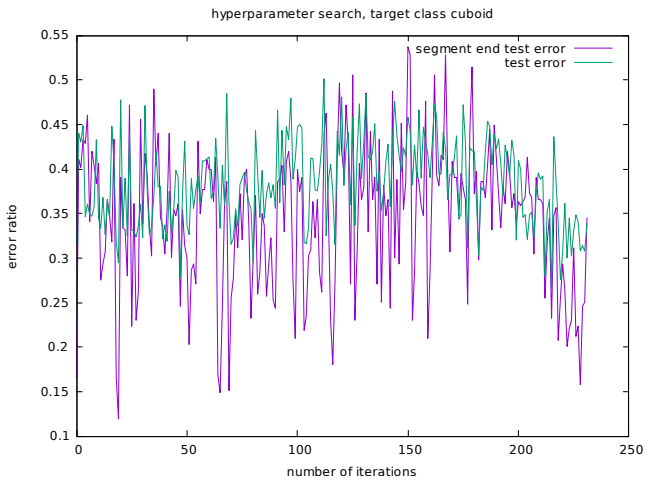


Fig. 10. Comparison between the test error and the test evaluated for the last 15% of points within a segment. Best classification accuracy is achieved for the number of neurons $N = 500$, leakage rate $\lambda = 0.1$, input scaling parameter equals 15.

hypothesis stating that haptic search, similar to visual search, is guided by a set of *haptic guiding features*. We have applied this hypothesis to create a target-distractor classifier for data acquired from human participants performing haptic search in a 3D environment. To address the main question posed in this paper, a suitable representation of haptic search data, we have compared a feed-forward network (ELM) and a recurrent network (ESN). The ESN yielded altogether better results than the ELM. The recurrent architecture of the ESN improved classification accuracy in comparison to the feed-forward ELM presumably through the temporal integration of the input. ESN yielded a higher classification accuracy for the multimodal input data compared to classification based on the individual modalities. Evaluation of the *end of segment* error yielded c.a. 90% classification accuracy, which is close to a human performance, but still demonstrates that we need a more sophisticated higher level tool to meaningfully accumulate the classifier output to make human-like decisions about moving to a new spot / identifying the object class.

To this end, we will extend our modeling approach in future work with a spatial representation of the acquired data to account for the movement of the hand in the environment. Along with better understanding human haptic search skills, further research will test whether the concept and representation presented in this paper will be able to autonomously guide a dexterous robot hand in the search of a target through the same stimuli as presented to the humans. The goal is to first perform these experiments in simulation, and then on a real robot hand equipped with sensors comparable to the tactile glove.

ACKNOWLEDGMENT

This work was supported by the DFG Center of Excellence EXC 277: Cognitive Interaction Technology (CITEC) and was partially funded from the EU FP7/2007-2013 project no. 601165 WEARHAP. We would like to thank Statex Production & Distribution plc. for providing conductive fabric samples for this project. We are also very grateful to Kathrin Krieger for help with the experimental execution.

REFERENCES

[1] J. M. Wolfe, "Guided search 4.0: Current progress with a model of visual search," *Integrated Models of Cognitive Systems*, 2007.

[2] R. Kõiva, T. Schwank, R. Haschke, and H. Ritter, "Fingernail with static and dynamic force sensing," 2016.

[3] G. Büscher, M. Meier, G. Walck, R. Haschke, and H. J. Ritter, "Augmenting curved robot surfaces with soft tactile skin," in *IROS*, Sept 2015, pp. 1514–1519.

[4] D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*. New York, NY, USA: Basic Books, Inc., 1979.

[5] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, 2010, visual Search and Selective Attention.

[6] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[7] S. R. Mitroff, A. T. Biggs, S. H. Adamo, E. W. Dowd, J. Winkle, and K. Clark, "What can 1 billion trials tell us about visual search?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 41, 2015.

[8] G. Bajlekov, "Theories of visual search and their applicability to haptic search," Master's thesis, Utrecht University, 2012.

[9] R. Martins, J. a. F. Ferreira, M. Castelo-Branco, and J. Dias, "Integration of touch attention mechanisms to improve the robotic haptic exploration of surfaces," *Neurocomput.*, 2017.

[10] V. S. Morash, "Detection radius modulates systematic strategies in unstructured haptic search." in *World Haptics*, 2015.

[11] K. Krieger, A. Moringen, R. Haschke, and H. Ritter, "Shape features of the search target modulate hand velocity, posture and pressure during haptic search in a 3d display," in *Lecture Notes in Computer Science*. Springer, 2016, both authors contributed equally to this equally.

[12] J. Nowainski, "Efficient target identification during haptic search in a three-dimensional environment," Bachelor Thesis, Bielefeld University, 2017, https://github.com/jnowainski/Bachelorarbeit/blob/master/tex/thesis/main.pdf.

[13] M. Grunwald, Ed., *Human Haptic Perception*. Birkhaeuser Verlag, 2008, ch. Haptic Shape Cues, Invariants, Priors, and Interface Design.

[14] K. Overvliet, K. Mayer, J. Smeets, and E. Brenner, "Haptic search is more efficient when the stimulus can be interpreted as consisting of fewer items," *Acta Psychologica*, vol. 127, no. 1, pp. 51 – 56, 2008.

[15] K. Overvliet, J. Smeets, and E. Brenner, "The use of proprioception and tactile information in haptic search," *Acta Psychologica*, vol. 129, no. 1, pp. 83 – 90, 2008.

[16] A. Moringen, R. Haschke, and H. Ritter, "Search procedures during haptic search in an unstructured 3d display," in *IEEE Haptics Symposium*, 2016.

[17] A. Moringen, K. Krieger, R. Haschke, and H. Ritter, "Haptic search for complex 3d shapes subject to geometric transformations or partial occlusion," in *IEEE World Haptics*, 2017.

[18] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *IEEE International Joint Conference on Neural Networks*, vol. 2, 2004, pp. 985–990.

[19] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *GMD Technical Report*, vol. 148, p. 34, 2001.

[20] J. Maycock, D. Dornbusch, C. Elbrechter, R. Haschke, T. Schack, and H. Ritter, "Approaching manual intelligence," *KI - Künstliche Intelligenz*, vol. 24, 2010.

[21] J. Maycock, T. Röhling, M. Schröder, M. Botsch, and H. Ritter, "Fully Automatic Optical Motion Tracking using an Inverse Kinematics Approach," in *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, 2015, pp. 461–466.

[22] G. H. Büscher, R. Kõiva, C. Schürmann, R. Haschke, and H. J. Ritter, "Flexible and stretchable fabric-based tactile sensor," *Robotics and Autonomous Systems*, vol. 63, pp. 244–252, 2015.

[23] W. Aswolinskiy, R. F. Reinhart, and J. Steil, "Time series classification in reservoir-and model-space," *Neural Processing Letters*, pp. 1–21, 2017.