Reach Prediction using Finger Motion Dynamics

Dimitar Valkov
Saarland University, DFKI, Saarland Informatics Campus
Saarbrücken, Germany
dimitar.valkov@dfki.de

Pascal Kockwelp
University of Münster
Münster, Germany
pascal.kockwelp@uni-muenster.de

Florian Daiber
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
florian.daiber@dfki.de

Antonio Krüger
DFKI, Saarland Informatics Campus
Saarbrücken, Germany
krueger@dfki.de

ABSTRACT

The ability to predict the object the user intends to grasp or to recognize the one she is already holding offers essential contextual information and may help to leverage the effects of point-to-point latency in interactive environments. This paper investigates the feasibility and accuracy of recognizing un-instrumented objects based on hand kinematics during reach-to-grasp and transport actions. In a data collection study, we recorded the hand motions of 16 participants while reaching out to grasp and then moving real and synthetic objects. Our results demonstrate that even a simple LSTM network can predict the time point at which the user grasps an object with 23 ms precision and the current distance to it with a precision better than 1 cm. The target's size can be determined in advance with an accuracy better than 97%. Our results have implications for designing adaptive and fine-grained interactive user interfaces in ubiquitous and mixed-reality environments.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; **Virtual reality**.

KEYWORDS

datasets, grasp prediction, hand gesture, neural networks

ACM Reference Format:
Dimitar Valkov, Pascal Kockwelp, Florian Daiber, and Antonio Krüger. 2023. Reach Prediction using Finger Motion Dynamics. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3544549.3585773

1 INTRODUCTION

The advent of consumer-grade virtual and augmented reality interfaces and powerful machine learning (ML) algorithms in the last decades have led to renewed interest in hand- and gesture-based natural user interfaces, where real or virtual objects can be directly grasped and manipulated. This is by no means surprising since our hands are remarkable instruments that help us shape, transform, use and manipulate our surroundings. In our everyday life, we use our hands for both - acting on *and* exploring the environment, as well as perceiving the shape, size, weight, surface texture, etc. of objects, through the sense of touch and proprioception. Consequently, researchers have approached the field from different directions with investigations spanning from advanced hand tracking [32] and gesture recognition [12] algorithms, through a diversity of approaches that detect touch and pressure with hand- or body-attached [10] sensors, to ingenious solutions to augment the surrounding with tunable haptics [15]. However, the unique ability of the hand to convey information about the environment and the user's perception of this environment has rarely been considered [33]. Indeed, the specific interrelation between the fingers that hold an object already contains information about the shape, size, and orientation of the object's graspable surfaces. More importantly, the gradual molding of the fingers during a reach-to-grasp (R2G) action has been shown to correlate with the object's physical properties [3], intended use [7], and the user's attitude toward that object [4]. This gradual molding is temporally and spatially correlated with the previously mentioned properties well before the hand actually reaches the intended object [20]. All these studies have one very important implication for interaction design: with sufficient environmental information and tracking fidelity, the black-box model of human prehension can be inverted. The correlation between objects and prehensile behavior indicates that the object, which the user intends to interact with, can be determined based solely on the temporal behavior of the fingers, captured with sufficient precision. Furthermore, the fact that grasp formation is recognizable *before* the hand actually reaches an object gives us the ability to *predict* the intended object. Unfortunately, while hand transport – governed by the Fitts' law – is a subject of detailed investigations for decades, prehensile kinematics has rarely been considered in the HCI domain.

In this paper, we address this challenge and have collected high-precision data for prehensile hand movements in a controlled data acquisition study. Our evaluations indicate that (a) both the current distance to the intended object and the moment at which the user grasps it can be predicted with high precision well before it is reached, and (b) the object can be discriminated during the R2G action, as well as when the user is already holding it if it has sufficiently distinguishable grasping affordances. These results have the potential to inform the design of future interactive environments, where virtual objects might be designed in a way that maximizes

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585773

their discriminability. Furthermore, this enables a more informed design of interfaces that make use of everyday objects as (usually mismatching) haptic proxies [31], and paves the way toward leveraging the effects of latency by using predictive algorithms able to discriminate future actions.

2 RELATED WORK

The high speed and dexterity of the hand make its precise tracking a challenging task for any tracking approach. Researchers have addressed the problem by developing complex ML [12], optimization [32], or rule-based [22] algorithms for hand pose reconstruction and gesture recognition [12]. However, most of these approaches are based on rigid initial assumptions about the correlations of the joint movements or learn these implicitly from hand motion datasets. This may lead to misinterpretations and data leakage when used in hand kinematics studies [19]. Similar to Heumer et al. [21] we based our evaluation on simple geometric properties extracted directly from the captured motion data. This has the additional benefit that similar features can be extracted with alternative, lightweight approaches, such as the Finexus [10], CyclopsRing [8], or FingerPad [9]. Finexus [10] in particular demonstrates how the position of the fingertips can be tracked in 3D space with high precision and speed, using only low-cost hardware.

Hand prehension is probably one of the most well-studied human activities in the various domains of psychology and neuroscience [1, 24]. In this context, we use the definition of *prehension* given in the seminal work of Mackenzie and Iberall, i.e., *"the application of functionally effective forces by the hand to an object for a task, given numerous constraints"* [24]. Interesting in this definition is, that it describes prehension by means of a cognitive black-box model, that converts the object's geometric properties and user's intent into a motor program steering the hand and finger motions [24]. Most of the work on hand prehension has focused on investigating the functionality of this black box. For instance, the biomechanical and kinematic properties of the hand and their effect on prehension have been investigated by Chen et al. [11] and Duncan et al. [17], some behavioral aspects were addressed e. g. by Jones & Lederman [1], and many neuro-psychological models of prehensile behavior have been proposed [16, 25, 29].

Research groups have also focused on studies on hand preshaping, in particular in the R2G task, and have consistently confirmed that grasp formation is highly correlated with the form and the size of the intended object (see e. g. [5] and [18] for recent surveys), as well as strongly related to the intended action [5, 18]. Furthermore, Chieffi et al. [13] and Ansuini [3] investigated the coordination between hand transport and grasp formation and showed that they are mostly independent but the hand pre-shaping is mostly finalized well before the hand reaches the object [13]. This observation is also confirmed by the work of Santello & Soechting [30], Molina-Vilaplana et al. [25] and some very preliminary evaluations in the HCI domain [14]. Unfortunately, in most cases, the captured motion sequences were trimmed, re-scaled to the uniform time interval $[0, 1]$, and equidistantly re-sampled. None of these steps can be performed in real-time since the entire sequence is required for each of them. Thus, it is currently not clear, how these results can be transferred to practical interactive systems. In the HCI domain, Paulson et al. [26] have investigated a grasp-shape-based selection of objects in office settings and Vatavu et al. [33] have developed a grasp-posture-based object recognizer for six basic geometric solids that achieved a recognition rate of about 60% in general, but up to 98% when using user-specific metrics. Xia et al. [34] have also demonstrated how a touch point can be predicted based on the trajectory of the hand in the ballistic phase during the hand transport task. Nevertheless, none of these works make use of the prehensile kinematics while the user is reaching out to grasp an object.

3 DATA COLLECTION

We collected high-precision hand and finger motion tracking data for R2G and object transport (OT) actions from 16 adults (13 self-identified as male, 3 as female, mean age $\mu = 26.13$, $\sigma = 1.78$)[1]. In this data collection study, the participants had to reach out with the right hand and grasp an object placed on a predefined *object position* (R2G task), and then lift the object and move it to a predefined *target position* (OT task), as illustrated in Figure 1a. We used 16 different objects and the task was repeated three times for each object, resulting in 48 data sequences per participant.

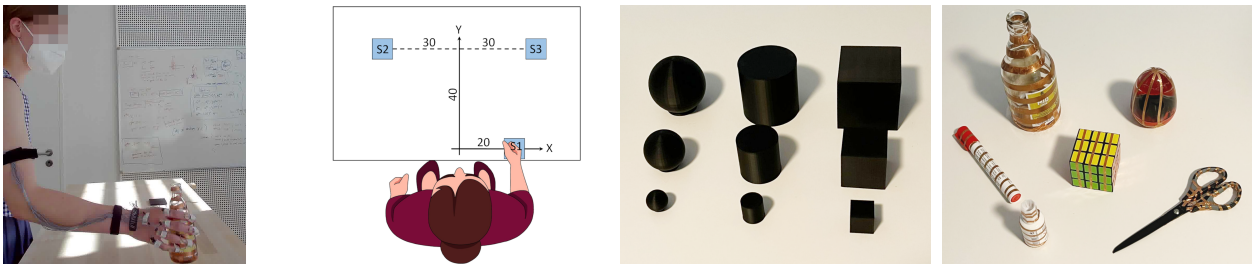
3.1 Setup

We used a word-space coordinate system with the xy-plane parallel to the participant's transverse plane and the yz-plane parallel to the sagittal plane (s. Figure 1b). The *hand resting position* was chosen slightly to the right for convenience, and the object and target positions were within the convenience range for all participants (Figure 1b). All three positions were implemented as touch-sensitive surfaces. These were 3D printed with conductive carbon-based filament (ProtoPasta CDP1175) and connected to a touch driver MPR121 and an Arduino Nano33 IoT board that captured the touch events at 250 fps.

The 16 study objects were split into two distinct sets - synthetic and "real" objects. The set of synthetic objects is shown in Figure 1c and consists of three regular geometric solids - sphere, box, and cylinder, each in three different sizes - small (2cm), medium (4cm) and large (6cm). The rationale behind the selected shapes is that we wanted to provoke the participants to use similar grasps for each object. For instance, a higher cylinder would have been easy to distinguish from a sphere since most participants would use a "cylindrical grasp" for the first and a "spherical grasp" for the second. With the selected objects we expected the participants to use the so-called precision grasp (thus to use the long fingers in opposition to the thumb, without hand opposition) in all cases with a different number of fingers, depending on the object's size. In this case, the discrimination algorithm will need to learn the very subtle kinematic differences in order to distinguish between objects of the same size with a different shape.

The set of real objects, illustrated in Figure 1d, consisted of seven objects one would commonly find in an office environment. We selected objects with appropriate size, that have well understandable grasping and usage patterns, and are handheld and commonly moved (e. g. not a puncher or a flowerpot). The resulting set (cf.

[1]The data set is available as supplementary material to this paper, and on the project website https://umtl.cs.uni-saarland.de/research/projects/grasp-prediction.html



(a) (b) (c) (d)

Figure 1: Experiment setup: (a) Test subject while participating in the experiment. (b) Experiment setup with the marked hand resting position and touch sensor $S1$, the object position and sensor $S2$, and the target position and sensor $S3$. A photograph of (c) the synthetic and (d) the real objects used in the experiment - *pen, glue, bottle, Rubik's cube, egg-vulcano, toy,* and *scissor.*

Figure 1d) is neither complete nor universally generalizable but sufficient to gain some initial insights into the domain.

We used a Polhemus Viper16 electromagnetic tracking system with 12 tethered micro sensors to track participant's hand and fingers with submillimeter precision at 960 fps. We attached 5 sensors to the fingernails and 5 sensors in the middle of the proximal phalanges of the fingers (cf. Figures 1 and 3). Two additional sensors were attached to the metacarpal of the thumb and to the metacarpal of the middle finger, respectively. The data capturing application was implemented in C/C++ using the Qt framework (v.6.3) and was run on a Windows 11 laptop with a Core i7-11800H processor, 32 GB RAM, and Nvidia RTX 3080 GPU. The synthetic objects were 3D printed with conductive filament and the real objects were enhanced with copper tape to enable touch transfer to the surfaces. The overall setup, although simple, enables high-fidelity motion capturing where each relevant event (e. g. when the participant lifted her hand from the start position or touched the object) can be reliably detected.

3.2 General Observations and Data Validity

From the 16 participants that took part in the experiment 768 tagged motion sequences were collected. From these, we excluded 5 paths from the evaluation because of erroneous (de-)activation of the touch surfaces or erroneous task performance. From the remaining 763 sequences, we extracted the R2G and the OT parts. Consistently

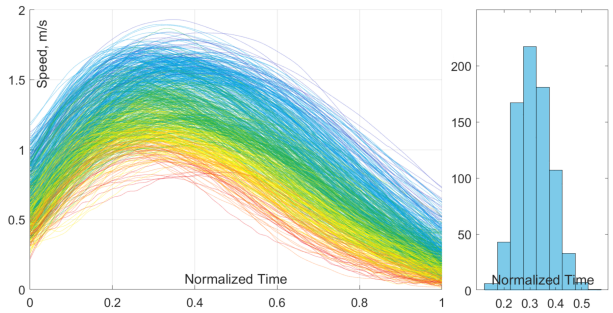


Figure 2: Speed profiles for all trials in the R2G phase (left), and histogram of the PMV in normalized time (right).

with the related work, e. g., [34], the motion profiles display clearly distinguishable acceleration and deceleration phases with a single (ignoring noise) maximum, called *point of maximum velocity (PMV)*, as illustrated in Figure 2. The PMV was mostly reached well before 50% of the motion but was not correlated with the absolute motion time ($R^2 = 0.07$).

In contrast to the related work [3], we did not control the initial hand shape, which eventually resulted in no clearly observable synchronization between hand transport and pre-shaping. Nevertheless, we observed that the change in the grip apertures becomes more systematic and stable as the hand approached the target. For all participants, this stabilization was well before the PMV. This is in partial correspondence with [3, 20], although we found no suitable method to quantify (and thus prove) these observations.

The execution time of the R2G phase varied between 365.7 and 1179.1 ms ($\mu = 645.64$ ms, $\sigma = 144.26$) and we found a significant main effect of *participant* ($F^{15}_{240} = 107.84$, $p < 0.01$), with four homogeneous groups (Scheffe's test $p > 0.98$ within the group, $p < 0.01$ between any group and non-group participant) of similarly performing participants. A repeated measurements ANOVA also revealed a significant within-subject effect of the object's size on the R2G execution time ($F^{2}_{18} = 10.94$, $p < 0.01$). Similarly, the execution time in the OT phase varied between 298.9 and 1350.0 ms ($\mu = 706.91$ ms, $\sigma = 185.59$) and we found a significant effect of *participant* ($F^{15}_{240} = 97.28$, $p < 0.01$), with three homogeneous groups of similarly performing participants. As expected, the grasp apertures of the fingers were (mostly) constant during this phase.

The effect of object size in the R2G phase is well expected and in accordance with the Fitts' law. In particular, participants reached for differently sized objects at the same distance, and their performance time increased for smaller objects. In contrast, the significant effect of participants in both phases is probably explained by the small sample size, and we don't expect it to generalize to larger populations. Overall, our general findings are consistent with and well explained by the current state of research, which motivates the overall validity of the dataset and provides some initial intuition for the methods used in the following sections.

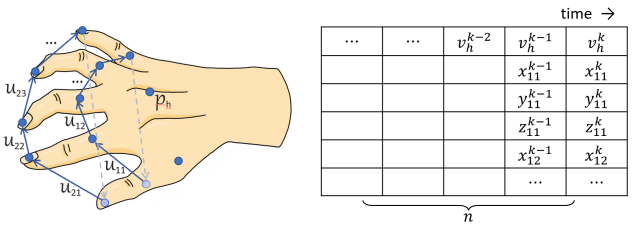


Figure 3: Illustration of the prediction model. The finger polygon vectors u_{ij} are calculated from the raw sensor positions and used together with the hand motion speed v_h to construct a feature sequence. Each table row represents a feature and each column represents a time sample.

4 PREDICTION MODEL

To extract more in-depth knowledge from the data set, we converted the raw sensor readings into higher-level feature sequences and used these to train long short-term memory neural networks (LSTM-NN) as illustrated in Figure 3. These networks were then used to predict the distance between the hand and the to-be-grasped object, the time until the grasp occurs, and the object's shape or size.

4.1 Feature Extraction

The thumb-to-index grip aperture is one of the most used approaches to describe a grasp pose [6]. An alternative is to use the surface, extend, or roundness of the (non-planar) polygon formed by the fingertips [23]. Inspired by these, we formulated a *finger polygon (FP) model* (see Figure 3 left) that describes the grip with two independent polygons and used the coordinates of the edge vectors $u_{ij} = (x_{ij}, y_{ij}, z_{ij})$ as features.

In addition, we used the (undirected) *hand movement velocity* $v_h(t)$ in our feature set. It was calculated using the finite difference approximation $v_h(t) = f \cdot \|p_h(t_k) - p_h(t_{k-1})\|$, where $p_h(t_k)$ denotes the hand position in frame k, and f is the tracker's frame rate. The velocity profile $v_h(t)$ was smoothed with a 1D median filter with kernel size 5 followed by an IIR low-pass filter, with a passband frequency of 25 Hz and steepness factor of 0.95.

The feature sets were then combined in short temporal sequences of either a fixed number of $n = 50$ samples (TM sequence) or a variable number of samples (PL sequence) as illustrated in Figure 3 right. We selected $n = 50$ (approx. 52 ms) for the PL sequence, as this will allow a barely noticeable initial response delay, even for short grasp gestures. The number of samples n in the PL sequences was calculated such that the path length s_h traveled by the hand is exactly

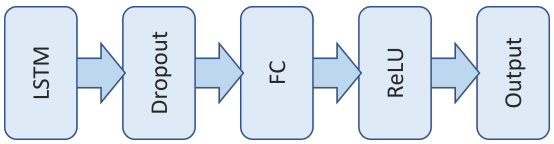


Figure 4: Architecture of the LSTM-NN used in our evaluations. The output layer is a single neuron with linear activation for the regression analysis, or a FC layer with *softmax* activation for the classification tasks.

2 cm long. We experimented with different path lengths for this subdivision scheme and found that shorter s_h yield generally better performance but are more susceptible to tracking noise. Thus we decided to use $s_h = 2$ cm as a thread-off between performance and robustness - this is well above the range of tracking jitter and hand tremor for healthy users. The intuition behind the PL subdivision is that a fixed-time segment would convey more information for a fast-performing participant compared to a slow-performing one. Thus, we hope that the PL sequences will normalize the information content of a single sequence across participants. The lengths of the PL sequences varied between 10 and 100 samples with a mean of $\mu = 24.11$ ($\sigma = 14.49$) for the R2G phase and $\mu = 24.26$ ($\sigma = 18.12$) for the OT phase.

The total number of data points for the R2G phase with PL subdivision was 26683, and with TM subdivision it was 11216. For the OT phase we had 31148 PL and 17066 TM data points.

4.2 Network Architecture

In this work we have selected to use LSTM-NN since they have already proven their utility for handling multivariate time series in numerous application fields [2, 27, 28]. The networks' architecture is depicted in Figure 4. The LSTM layer used a *tanh* state and *sigmoid* gate activation functions and was linearly connected to the subsequent fully connected layer FC. To avoid overfitting, a dropout layer with a rate of 0.2 was connected between LSTM and FC layers. The output layer consisted of a single fully connected neuron with linear activation for the regression networks, or of a fully connected layer with *softmax* activation for the classification networks. To select the sizes of the LSTM and FC layers, we started each evaluation with a small network and increased progressively the numbers of (hidden) neurons to find the best performance while still avoiding overfitting. The concrete values are presented in the respective sections.

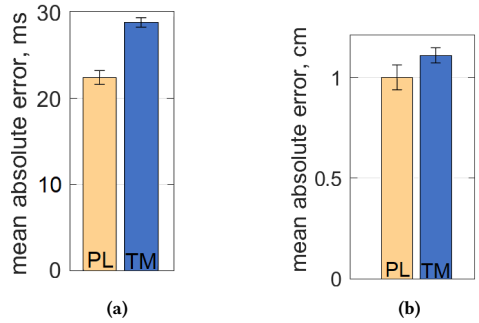
In all cases, the training sets were shuffled once at the beginning, and the networks were trained with the Adam optimizer using L2 norm gradient threshold, L2 regularization with $\alpha = 10^{-4}$, and mini-batch size of 64. We used the root mean square error as a loss function in the regression, and the cross-entropy loss in the classification tasks.

5 RESULTS

5.1 Time until Grasp

For this evaluation, we calculated the time interval between the current sample and the end of the R2G phase and trained the regression LSTM-NN to predict it. The network was configured with 64 hidden neurons in the LSTM layer and 16 neurons in the FC layer and was trained for 100 epochs with a learning rate $\eta = 1 \cdot 10^{-3}$. Essentially, this is aimed to simulate a run-time system that uses the latest n tracking samples to predict the time until the user grabs an object.

Figure 5a shows the mean average error (MAE) for the network trained with the PL and TM sequences (4-fold factored cross-validation, repeated twice). As illustrated in the figure, the training and validation were mostly consistent for each fold and repetition and the PL sequence achieved a better overall accuracy ($\mu = 22.38$



(a) (b)

Figure 5: Accuracy of the LSTM-NN when trained to predict (a) the time until the user reaches the object and (b) the distance to the object.

ms, $\sigma = 0.792$ ms) compared to the TM sequence ($\mu = 28.56$ ms, $\sigma = 0.523$ ms).

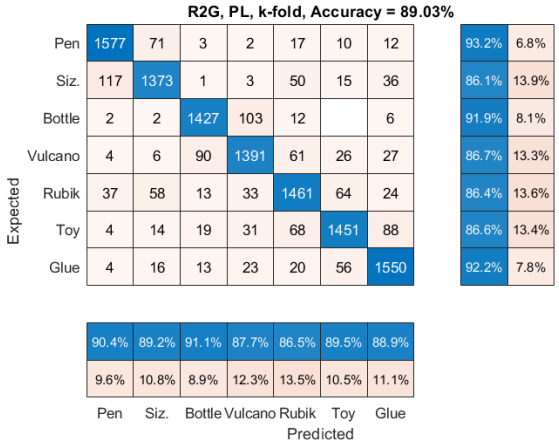
To test whether the network can generalize over unknown users we conducted a leave-one-user-out test. In this test, the LSTM was trained with the PL sequences of all but one user and tested with the data of the left-out user[2]. The average performance dropped to $\mu = 68.12$ ms ($\sigma = 25.72$ ms) with a more noticeable performance loss for TP16 (MAE = 129.14 ms), TP20 (MAE = 109.0 ms), and TP4 (MAE = 102.36). We have carefully investigated the experiment video recordings for these three participants (2 male and 1 female, all right-handed) but found no obvious differences compared to the remaining test population. Notably, even the worst results (MAE $\leq$ 130 ms) are still considerably better than a chance (MAE $\approx$ 400 ms) or a "loose-fit" regression with $R^2 = 0.5$ (MAE $\approx$ 200 ms). Thus, the networks can transfer what they learned to unknown users and make a reasonable, albeit sometimes a bit unprecise, estimation about the expected time until the user reaches the intended object.

5.2 Distance to Target

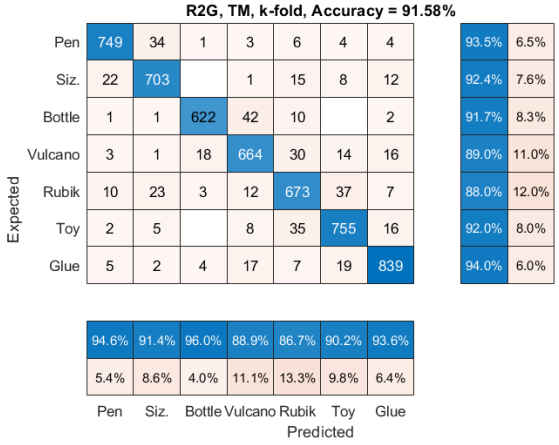
In this evaluation, the regression LSTM-NN, with the same configuration as in the previous section, was trained to predict the distance between the current hand position and the object position in the R2G phase.

Figure 5b plots the results from the 4-fold cross-validation (repeated twice) for the two sequence types. Again, the networks showed consistent performance in each fold and repetition with PL ($\mu = 1.00$ cm, $\sigma = 0.062$ cm) outperforming the TM ($\mu = 1.11$ cm, $\sigma = 0.037$ cm) sequences. The MAE for the leave-one-out validation with the PL sequence ranged between 1.04 and 2.98 cm ($\mu = 1.70$ cm, $\sigma = 0.546$ cm). Thus, even the worse accuracy values are still well within the useful range for practical application. Furthermore, the network's precision increases as the hand approaches the target position, with MAE = 0.736 cm when the hand is within the last 20 cm, MAE = 0.917 cm for the interval 20 – 40 cm, and MAE = 1.128 cm when the hand is more than 40 cm away from the target.

[2]We also conducted leave-one-session-out and leave-one-object-out (for the synthetic objects) tests but found only a marginal drop in the overall accuracy, with no apparent influence of the left-out object or session on the performance. Thus, we omit these here and in the following sections for briefness.



(a)



(b)

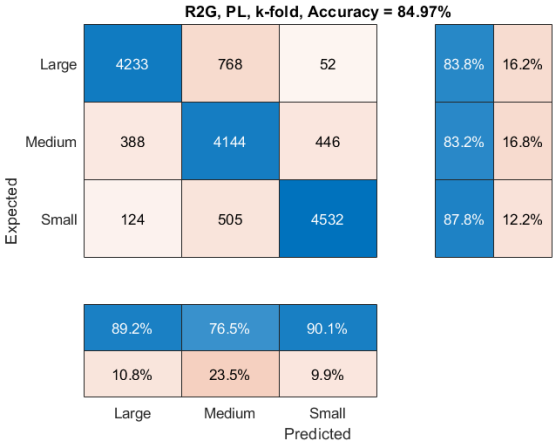
Figure 6: Confusion matrices for the discrimination of real objects in the R2G phase for (a) PL and (b) TM sequences.

5.3 Object Discrimination during R2G

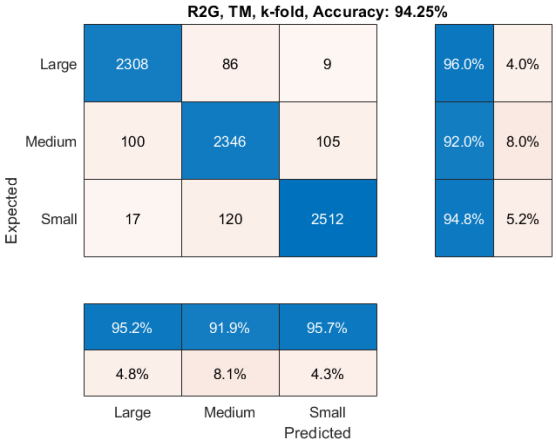
In this evaluation, we tested the network's ability to recognize the object the user will grasp. As described previously, we used two distinct sets of objects - *real* and *synthetic*. For the synthetic objects, we further split the evaluation into two parts - recognition of the object's *size* and *shape*. The network was configured with 128 hidden neurons in the LSTM layer and 16 neurons in the FC layer.

Figure 6 presents the confusion matrices and recognition accuracy for the real objects with the networks trained with the PL and TM sequences and Figure 7 depicts the performance for the *size* of the synthetic objects.

When trained to discriminate the object's *shape* for the synthetic objects (i. e. box, cylinder, or sphere) the networks achieved a high accuracy (better than 80%) with the k-fold cross-validation, but it dropped to almost a chance level in the leave-one-out tests, which is

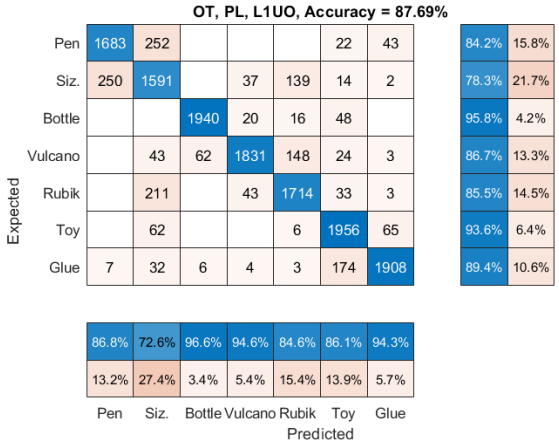


(a)

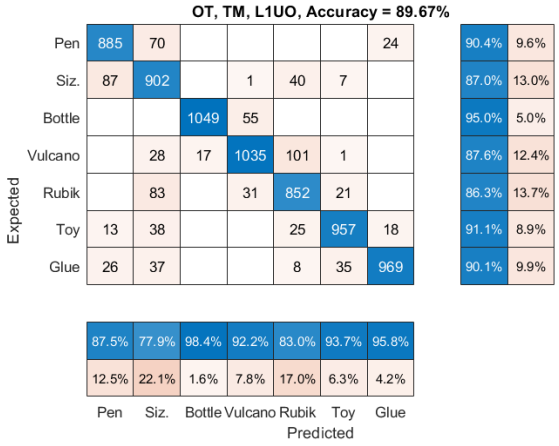


(b)

Figure 7: Confusion matrices for the size discrimination of the synthetic objects in the R2G phase for (a) PL and (b) TM sequences.



(a)



(b)

Figure 8: Confusion matrices for the discrimination of real objects in the OT phase for (a) PL and (b) TM sequences.

a clear indication that the algorithm rather overfitted in the training data instead of learning to extract useful information. In contrast, the recognition of the object's size remained robust in the leave-one-out test with an accuracy of 67.88% for the PL sequences, and an accuracy of 73.63% for the TM sequences. For the real objects, the recognition accuracy dropped to 44.77% for the PL sequences, and to 55.44% for the TM sequences. Nevertheless, in both cases, these accuracy levels are significantly better than the chance level and the confusion matrices depict a clear correlation between the real and the predicted classes.

As one can see in Figure 6, the algorithms have some problems distinguishing between the *toy* and the *glue*, both grasped sidewards with a three- or four-finger pinch grip. Interestingly, some participants grasped the *vulcano* object with a spherical grip, while others used a sidewards precision grip. Therefore, it was commonly confused with either the *Rubik's cube* or the *bottle* (depending on the participant) by the recognizer. Similarly, the *scissors* (abbreviated *siz.* in the Figure) was confused with either the *pen* when grasped in the middle with a three-finger pinch grasp from above or with the *Rubik's cube* when the participant grasped it with a wider grip on the handles. The behavior is qualitatively similar for both the TM and the PL sequences, which indicates that it is due to the intrinsic complexity of the overall problem.

As expected, the size recognition task (see Figure 7) was more successful, with the medium-sized objects equiprobably confused with small or large objects.

5.4 Object Discrimination during OT

In the last evaluation, we tested the network's ability to recognize the object the user is already holding. This might be useful for interfaces that use everyday objects as haptic or interaction proxies and in the context of AR-enhanced ubiquitous computing. For the evaluation, we used the same network as in the previous section,

testing the discriminability of the intended object (for the real objects) or the object's size and shape (for the synthetic objects).

The shape discrimination for the synthetic object failed again to yield trustful results for all tested combinations of features and hyperparameters. In contrast, the size discrimination task seems to be considerably less complicated in this phase, and the k-fold cross-validation test yielded accuracy levels better than 99% in all cases. The network trained with the TM sequences also performed better in the leave-one-out test, achieving a mean accuracy of $\mu = 98.58\%$ ($\sigma = 2.40\%$) with a 100% accuracy for 12 out of the 16 participants and worst performance of 95.89%.

Similar behavior was observed in the object discrimination task for the real objects, as illustrated in Figure 8. In the k-fold cross-validation, both the TM and PL sequences achieved an accuracy better than 99%, and in the leave-one-out test, the TM sequence outperformed the PL with a mean accuracy of $\mu = 89.67\%$ ($\sigma = 9.02\%$). Again, objects with similar grasping affordances were commonly confused by the recognizer, albeit less severe in this phase.

6 DISCUSSION AND FUTURE WORK

The results presented in this work demonstrate how one can determine with high precision the distance to the to-be-grasped object, the approximate time until it is grasped, and at least the object size, using only local hand dynamics. The presented algorithms also showed promising results for discriminating objects with sufficiently different grasping affordances. This makes a tremendous amount of additional information available to the interface application, long before the user ever reaches an object. Indeed, even the fastest participant needed approximately 300 ms between the PMV and the first touch of the object. The two regression networks were able to deliver high-precision predictions from the very first feature sequence. Thus, even in this exceptional case, the interface will still have more than 280 ms to properly use the information. In the general case, this time buffer increases to more than 400 ms.

The time until grasp could be reliably estimated with an accuracy within the synchronization precision of the dataset (approx. 25 ms), and the accuracy of the distance estimations reached MAE of 1 cm for known and just under 3 cm (worst case) for unknown users. With the target being 64 cm away from the hand's starting position, a mean error of just a centimeter is remarkable. Moreover, these accuracy levels were achieved despite the significant difference in users' performance and grasping behavior. It is worth also repeating that the features were extracted directly from the marker data, without any user-specific adjustments, e. g., for different finger sizes.

A straightforward approach to increase the performance of the regression networks even further would be to use a (smart) post-processing filter for the estimations. The feature extraction presented in Section 4.1 allow generating a new prediction for each new tracking sample. We can conservatively expect that the newly predicted time-until-grasp will decrease by $0 - 1.1$ ms and the distance to the target will decrease by $0 - v_h \cdot 1.1$ ms for each new sample. We already successfully experimented with simple filtering of the form $p = k \cdot p_{\text{old}} - (1-k) \cdot p_{\text{new}}$, where p_{old} and p_{new} are the old and the new predictions, and k is selected from a probability distribution over the conservative intervals. Nevertheless, a new dataset will be needed for further evaluations in this direction in order to avoid over-optimization.

Disappointingly, the object's shape could not be reliably discriminated for the synthetic objects with our algorithms. The accuracy changed considerably for the real objects, which have clearly distinguishable grasp affordances, but it is still below the level needed for usable applications. With the standard k-fold cross-validation we did achieve recognition rates that are comparable to or even better than those reported in the related work [21, 33, 34]. Nevertheless, the leave-one-out tests showed that these do not generalize to unknown users and in some cases even to new trials from known users. Unfortunately, we did not measure the participants' hand and finger sizes and could not test whether these explain the low performance in the leave-one-out test. Nevertheless, we believe that the grasp type is a considerably stronger factor in this regard. The shape recognition of the synthetic objects makes this particularly evident. In this case, all objects were deliberately selected with similar grasping affordances, which made them indistinguishable for the algorithm. The common confusion of the similarly shaped real objects, e.g., glue and toy, is another indication of this. Indeed, with our approach, the network needed to recognize not only the target object itself but also *where* and *how* the user will grasp it while abstracting from human factors at the same time. In contrast, the size discrimination networks could safely rely on the number of folding fingers as a strong predictor and achieved high recognition rates for both the R2G and OT phases.

One possible solution would be to split the object recognition task into two phases: (a) prediction of the grasp shape and size and (b) mapping the grip to a target object. As already demonstrated, the size can be reliably predicted. Based on the results for the real objects, early prediction of the (class of the) grasp also seems plausible. Nevertheless, herefore we need again a new dataset in which the objects are selected w.r.t all common grasp types, which will be the subject of future experiments.

Another promising direction for future work is the definition of feature sets that are less dependent on hand or finger sizes. Trivial candidates are the finger joint angles or the distances between the fingertips. However, it is currently difficult to predict their utility, especially if they are calculated from (mathematically) limited sensor data.

Our initial intuition that the variable length sequences PL would equalize the inter-subject variations was only correct in the regression analyses, where PL significantly outperformed the TM sequences. In contrast, size and (real) object type discriminators clearly benefited from the constant length of the TM sequences.

Overall, the results already provide a solid working base for practical applications. For instance, the target distance prediction enables a rough estimation of the object's position. While we did not use the global hand position, it will certainly be known to the interaction system. Thus, using the moving hand position as a center and the estimated (undirected) distance as a radius will allow us to find the approximate object position as an intersection of spheres. The additional information about the object's size can further reduce the number of potential target candidates. A virtual environment – fully controlled by the designer – can be constructed such that this information is already sufficient. For instance, one can group items with different sizes and grasp affordances together while keeping

similarly sized objects further apart. The time estimations, albeit unprecise for some users, can further enable dynamic and adaptive interfaces. Nevertheless, the presented results are still preliminary and unveil a number of challenges that will be addressed in future research.

ACKNOWLEDGMENTS

We want to thank Valeria for helping design and shape this paper and the reviewers for their valuable comments and suggestions. This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project 436291335 and is part of Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments.

REFERENCES

[1] L A Jones and S J Lederman. 2006. *Human Hand Function*. Vol. 32. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195173154.001.0001

[2] Florent Altché and Arnaud de La Fortelle. 2018. An LSTM Network for Highway Trajectory Prediction. *CoRR* abs/1801.07962 (2018), 1–10. arXiv:1801.07962 http://arxiv.org/abs/1801.07962

[3] Caterina Ansuini, Andrea Cavallo, Atesh Koul, Marco Jacono, Yuan Yang, and Cristina Becchio. 2015. Predicting Object Size from Hand Kinematics: A Temporal Perspective. *PLOS ONE* 10, 3 (03 2015), 1–13. https://doi.org/10.1371/journal.pone.0120432

[4] Caterina Ansuini, Marco Santello, Stefano Massaccesi, and Umberto Castiello. 2006. Effects of end-goal on hand shaping. *Journal of neurophysiology* 95, 4 (2006), 2456–2465. https://doi.org/10.1152/jn.01107.2005

[5] Sonia Betti, Giovanni Zani, Silvia Guerra, Umberto Castiello, and Luisa Sartori. 2018. Reach-To-Grasp Movements: A Multimodal Techniques Study. *Frontiers in Psychology* 9, 2 (2018), 990.

[6] Umberto Castiello. 2005. The neuroscience of grasping. *Nature reviews. Neuroscience* 6 (10 2005), 726–36. https://doi.org/10.1038/nrn1744

[7] Andrea Cavallo, Atesh Koul, Caterina Ansuini, Francesca Capozzi, and Cristina Becchio. 2016. Decoding intentions from movement kinematics. *Scientific Reports* 6 (2016), 37036 EP –. https://doi.org/10.1038/srep37036

[8] Liwei Chan, Yi-Ling Chen, Chi-Hao Hsieh, Rong-Hao Liang, and Bing-Yu Chen. 2015. CyclopsRing: Enabling Whole-Hand and Context-Aware Interactions Through a Fisheye Ring. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 549–556. https://doi.org/10.1145/2807442.2807450

[9] Liwei Chan, Rong-Hao Liang, Ming-Chang Tsai, Kai-Yin Cheng, Chao-Huai Su, Mike Y. Chen, Wen-Huang Cheng, and Bing-Yu Chen. 2013. FingerPad: Private and Subtle Interaction Using Fingertips. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 255–260. https://doi.org/10.1145/2501988.2502016

[10] Ke-Yu Chen, Shwetak N. Patel, and Sean Keller. 2016. Finexus: Tracking Precise Motions of Multiple Fingertips Using Magnetic Sensing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1504–1514. https://doi.org/10.1145/2858036.2858125

[11] F. Chen Chen, A. Favetto, M. Mousavi, E. P. Ambrosio, S. Appendino, Battezzato A., D. Manfredi, F. Pescarmona, and B. Bona. 2011. Human Hand: Kinematics, Statics and Dynamics. In *International Conference on Environmental Systems*. AIAA, Portland, Oregon, 1–10. https://iris.polito.it/handle/11583/2460637

[12] H. Cheng, L. Yang, and Z. Liu. 2016. Survey on 3D Hand Gesture Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 9 (Sep. 2016), 1659–1673. https://doi.org/10.1109/TCSVT.2015.2469551

[13] Sergio Chieffi and Maurizio Gentilucci. 1993. Coordination between the Transport and the Grasp Components During Prehension Movements. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale* 94 (02 1993), 471–7. https://doi.org/10.1007/BF00230205

[14] Florian Daiber, Dimitar Valkov, Frank Steinicke, Klaus H. Hinrichs, and Antonio Krüger. 2012. Towards Object Prediction based on Hand Postures for Reach to Grasp Interaction. In *Proceedings of the ACM CHI Workshop on The 3rd Dimension of CHI: Touching and Designing 3D User Interfaces (3DCHI) 2012*. ACM.

[15] Donald Degraen, André Zenner, and Antonio Krüger. 2019. Enhancing Texture Perception in Virtual Reality Using 3D-Printed Hair Structures. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300479

[16] Cosimo Della Santina, Matteo Bianchi, Giuseppe Averta, Simone Ciotti, Visar Arapi, Simone Fani, Edoardo Battaglia, Manuel Giuseppe Catalano, Marco Santello, and Antonio Bicchi. 2017. Postural Hand Synergies during Environmental Constraint Exploitation. *Frontiers in Neurorobotics* 11 (08 2017). https://doi.org/10.3389/fnbot.2017.00041

[17] Scott F.M. Duncan, Caitlin E. Saracevic, and Ryosuke Kakinoki. 2013. Biomechanics of the Hand. *Hand Clinics* 29, 4 (nov 2013), 483–492. https://doi.org/10.1016/j.hcl.2013.08.003

[18] Ida Egmose and Simo Køppe. 2018. Shaping of Reach-to-Grasp Kinematics by Intentions: A Meta-Analysis. *Journal of Motor Behavior* 50, 2 (2018), 155–165. https://doi.org/10.1080/00222895.2017.1327407 arXiv:https://doi.org/10.1080/00222895.2017.1327407 PMID: 28644719.

[19] Q. Fu and M. Santello. 2011. Towards a complete description of grasping kinematics: A framework for quantifying human grasping and manipulation. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 8247–8250. https://doi.org/10.1109/IEMBS.2011.6092033

[20] M.P. Furmanek, L.F. Schettino, and M. Yarossi. 2019. Coordination of reach-to-grasp in physical and haptic-free virtual environments. *Journal of NeuroEngineering Rehabilitation* 16, 78 (2019). https://doi.org/10.1186/s12984-019-0525-9

[21] Guido Heumer, Heni Ben Amor, and Bernhard Jung. 2008. Grasp Recognition for Uncalibrated Data Gloves: A Machine Learning Approach. *Presence* 17, 2 (2008), 121–142. https://doi.org/10.1162/pres.17.2.121

[22] M. Höll, M. Oberweger, C. Arth, and V. Lepetit. 2018. Efficient Physics-Based Implementation for Realistic Hand-Object Interaction in Virtual Reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 175–182. https://doi.org/10.1109/VR.2018.8448284

[23] Sing Bing Kang and K. Ikeuchi. 1994. Determination of motion breakpoints in a task sequence from human hand motion. In *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*. 551–556 vol.1. https://doi.org/10.1109/ROBOT.1994.351241

[24] Christine L MacKenzie and Thea Iberall. 1994. *The grasping hand*. Amsterdam ; New York : North-Holland.

[25] Javier Molina-Vilaplana, Jorge Feliú Batlle, and Juan López Coronado. 2002. A Neural Model of Spatio Temporal Coordination in Prehension. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN '02)*. Springer-Verlag, London, UK, UK, 9–14. http://dl.acm.org/citation.cfm?id=646259.758872

[26] B. Paulson, D. Cummings, and T. Hammond. 2011. Object interaction detection using hand posture cues in an office setting. *International Journal of Human Computer Studies* 69, 1-2 (2011), 19–29. https://doi.org/10.1016/j.ijhcs.2010.09.003

[27] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics* 69 (2017), 218–229. https://doi.org/10.1016/j.jbi.2017.04.001

[28] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *CoRR* abs/1402.1128 (2014), 1–10. arXiv:1402.1128 http://arxiv.org/abs/1402.1128

[29] Marco Santello, Matteo Bianchi, Marco Gabiccini, Emiliano Ricciardi, Gionata Salvietti, Domenico Prattichizzo, Marc Ernst, Alessandro Moscatelli, Henrik Jörntell, Astrid M.L. Kappers, Kostas Kyriakopoulos, Alin Albu-Schäffer, Claudio Castellini, and Antonio Bicchi. 2016. Hand synergies: Integration of robotics and neuroscience for understanding the control of biological and artificial hands. *Physics of Life Reviews* 17 (2016), 1 – 23. https://doi.org/10.1016/j.plrev.2016.02.001

[30] M. Santello and J. F. Soechting. 1998. Gradual molding of the hand to object contours. *Journal of neurophysiology* 79, 3 (1998), 1307–1320. https://doi.org/10.1152/jn.1998.79.3.1307

[31] Adalberto L. Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality: Using the Physical Environment to Design Virtual Reality Experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. https://doi.org/10.1145/2702123.2702389

[32] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. 2016. Efficient and Precise Interactive Hand Tracking Through Joint, Continuous Optimization of Pose and Correspondences. *ACM Trans. Graph.* 35, 4, Article 143 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925965

[33] Radu-Daniel Vatavu and Ionu Alexandru Zaii. 2013. Automatic Recognition of Object Size and Shape via User-dependent Measurements of the Grasping Hand. *Int. J. Hum.-Comput. Stud.* 71, 5 (May 2013), 590–607. https://doi.org/10.1016/j.ijhcs.2013.01.002

[34] Haijun Xia, Ricardo Jota, Benjamin McCanny, Zhe Yu, Clifton Forlines, Karan Singh, and Daniel Wigdor. 2014. Zero-Latency Tapping: Using Hover Information to Predict Touch Locations and Eliminate Touchdown Latency. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 205–214. https://doi.org/10.1145/2642918.2647348