The Road Towards Image-Computable Models of Human Visual Grasp Planning

Guido Maiello[1], Lina K. Klein[1], Vivian C. Paulun[1], Katherine R. Storrs[1], Roland W. Fleming[1]

[1] Department of Experimental Psychology, University of Gießen, Germany

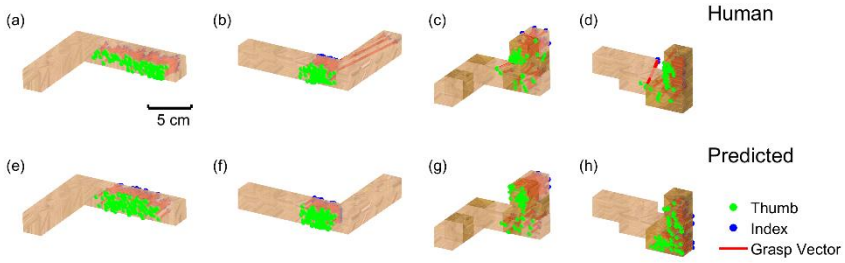

Figure 1. (a-d) Human grasps onto objects varying in shape, orientation, and material (wood and brass). (e-h) Simulated grasps onto the same objects, predicted through our theoretical framework.

Humans predominantly use vision to plan actions towards objects. Glancing at a nearby object, reaching out, and grasping it, feels effortless. However, the sensorimotor computations underlying grasp planning are nontrivial, and there is an extensive literature describing the multifaceted features of visually guided grasping[1,2]. At last year's meeting[3], we presented preliminary work aimed at predicting how humans visually select grasp locations on 3D objects. Since then, we have developed this work into a theoretical framework that unifies the varied yet fragmented literature on human grasp selection. We are now able to generate predictions of two-digit precision grip grasps onto 3D objects varying in shape, weight, and material (Figure 1, e-h). These predictions are strikingly similar to real grasps executed by human participants (Figure 1, a-d; behavioural data from [4]).

To generate these compelling predictions, we first create a triangulated 3D mesh model of a graspable object, and place it within a 3D coordinate frame. Within the same coordinate frame, we define the position and orientation of a human observer poised to grasp the object. We then sample the surface of the mesh model in discrete steps. Each sampled point on the surface of the object represents a potential contact location between the object and the hand grasping it. In precision grip, only two contact points are employed: the thumb contact point and the index finger contact point. Every possible precision grip grasp onto the object is thus defined as a 6D vector of x,y,z coordinates for the thumb contact point and x,y,z coordinates for the index contact point.

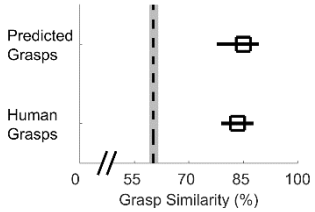

Figure 2. Similarity to the median human grasp. Dashed line is the chance level of similarity between grasps constrained by object geometry.

For each grasp within the 6D manifold defined by the surface of the object, we compute how far the grasp is from a set of optimality criteria which we hypothesize the brain may employ to plan a successful grasp. We then assign a set of penalty values to each grasp, proportional to the distance of the grasp from this set of optimality criteria. We include criteria determined by the physical properties of both the graspable object and the human actuator (i.e. the human arm/hand). Specifically, we consider grasp optimality criteria based on (i) optimum force closure[5], (ii) minimum torque[6], (iii) alignment with the natural grasp axis[7], (iv) optimal grasp aperture[8], and (v) minimum reach trajectory[9]. We assume that humans will most likely select a grasp that satisfies all optimality criteria as much as possible. This is equivalent to searching for a grasp that has low penalty across all optimality criteria. The predicted grasps in Figure 1 (e-h) are sampled from the set of grasps that best satisfy this intersection of constraints.

To quantitatively assess how similar predicted grasps are to human grasps, we defined a grasp similarity metric inversely proportional to the Euclidean distance (in 6D) between grasps, expressed as a percentage of the maximum possible distance between grips for each object. By this metric, Figure 2 shows how grasps selected by different human participants on the same object are 83.5% similar to each other, and grasps predicted through our framework are 85.1% similar to human grasps. With no direct knowledge of human data (i.e. without fitting human grasp data to a model), our approach can predict human grasps equally well as grasps from a random human on average approximate the median human grasp.

The simple, equally weighted intersection of the constraints we have thus far described already well predicts human grasping behaviour. However, it is unlikely that all optimality criteria will be treated as being equally important by humans selecting grasps on different objects. Additionally, different persons may give different relative importance to different constraints. Therefore, we developed a method of varying the importance of each optimality criterion to fit the intersection of the constraints to observed patterns of human behaviour. Through this procedure, we demonstrate that the combination of force closure, hand posture, and grasp size explains most of human grasp selection. The length of the reach trajectory marginally influences human grasp planning, and only for very light objects. Furthermore, we find that humans select grasps that minimize torque only on heavy objects where very high torques may occur. In addition to describing individual patterns of human grasps, our framework can be employed to generate novel, perceptually dissimilar stimuli, that differentiate conflicting hypotheses on how humans grasp objects (Figure 3).

Figure 3. Novel stimuli generated to differentiate conflicting hypotheses on visually guided grasping behavior. These objects are selected to be equally and maximally distinct, as determined by a shape dissimilarity metric we have perceptually validated in a visual similarity task.

We have thus far developed a framework capable of identifying the computations necessary to plan successful grasps when total knowledge of the physics of the environment is available. Hand-engineering algorithms that perform these computations directly from image input is a daunting task, one which has stumped the robotics community for decades[10]. Yet advances in machine learning may now come to our aid. Recently, deep convolutional neural networks have been successfully trained to control robotic grippers from monocular images[11]. Similarly, we are now attempting to train convolutional networks on motor tasks that a priori seem to require detailed physical knowledge of the world. Preliminary attempts suggest that genetic algorithms may be better than reinforcement learning for training deep convolutional neural networks at complex motor tasks[12]. This approach will hopefully allow us to determine the visual representations necessary for successful mappings between visual input and motor actions. Once we identify the visual computations that lead to successful motor actions, we will determine which of these computations the human visual system is employing to plan grasps towards objects. To this end, we are designing a dataset that will hopefully function as a benchmark to assess models of human grasping behaviour. At the end of my talk, I will present a preliminary version of this dataset, and I will seek feedback from the MODVIS community to help determine the structure and content of the dataset and ensure its success. Through a coordinated effort, the theory- and data-driven approach we present here holds the potential for developing complete, image-computable models of human visually guided grasping behaviour.

References:[1] Cuijpers RH, Smeets JBJ, Brenner E (2004) *J Neurophysiol* 91(6),2598-606 [2] Kleinholdermann U, Franz VH, Gegenfurtner KR (2013) *J Vis* 13(8), 23 [3] Maiello G, Klein LK, Paulun VC, Fleming RW (2017) *Modvis* [4] Klein LK, Maiello G, Paulun VC, Fleming RW (in preparation) [5] Nguyen VD (1988) *Int J Rob Res* 7(3), 3-16 [6] Lukos J, Ansuini C, Santello M (2007) *J Neurosci* 27(14), 3894-3903 [7] Schot WD, Brenner E, Smeets JB (2010) *Exp Brain Res* 204(2), 163-171 [8] Cesari P, Newell KM (1999) *J Exp Psychol Hum Percept Perform*, 25(4), 927 [9] Paulun VC, Kleinholdermann U, Gegenfurtner KR, Smeets JB, Brenner E (2014) *Exp Brain Res* 232(7), 2061-2072 [10] Saxena A, Driemeyer J, Ng AY (2008) *Int J Rob Res* 27(2), 157-173 [11] Levine S, Pastor P, Krizhevsky A, Ibarz J, & Quillen D (2017) *Int J Rob Res* [12] Such FP, Madhavan V, Conti E, Lehman J, Stanley KO, Clune J (2017) *arXiv*:1712.06567