State of the Art on: Perception for Robotic Grasping

Luca Cavalli, luca3.cavalli@mail.polimi.it

1. Introduction

Research in Perception for Robotic Grasping has the objective of providing new methods for the perception and environment modeling in Robotic Grasping. The Robotic Grasping research field tackles the problem of autonomously planning and executing effective grasps on objects. Robotic Grasping is tightly connected with the broader field of affordance perception, on whose definition we will be more precise in Section 1.2. The core community is in the Robotics research area as main field of application, but connections exist also with the areas of Computer Vision, Machine Learning and Artificial Intelligence in general as they provide fundamental tools to tackle this problem. Analyzing the affordance perception research community we can also find significant connections with the area of Human Computer Interaction with the aim of studying effective human-robot collaboration, with a particular focus on language-based communication.

The main venues interested in research on Perception for Robotic Grasping are the ones dealing with robotics and computer vision. Tables 1 and 3 show the selected top venues for Scientific Robotics, while Table 2 shows a selection of Computer Vision conferences. Conferences have been selected by H5-index, number of received submissions, rate of accepted submissions, review process and reviewers base, while journal selection takes into consideration H-index, foundation year and rate of articles remained uncited after one year from publication. The recent interest in the field of affordance perception is also witnessed by the organization of dedicated workshops; the latests include *Learning Object Affordances: a fundamental step to allow prediction, planning and tool use?* at IROS 2015 and the *International Workshop on Computational Models of Affordances in Robotics* at RSS 2018.

Conference Name	H5-index	Yearly submissions	Acceptance Rate	Review process	Reviewers base
ICRA	75	around 2000	40-45%	single blind peer review	thousands
IROS	54	1500-2500	30-50%	single blind peer review	hundreds
RSS	49	150-250	20-40%	double blind peer review	tens

Table 1: Top conferences in Scientific Robotics

Conference Name	H5-index	Yearly submissions	Acceptance Rate	Review process
CVPR	158	around 3000	30%	double blind peer review
ECCV	98	around 1500	25-30%	double blind peer review
ICCV	89	around 1200	25-30%	double blind peer review

Table 2: Top conferences in Computer Vision

1.1. Preliminaries

Fundamental tools for the grasping scientific research community come from physics, and they are used to quantitatively evaluate the quality of a grasp given the hand model, its configuration, and contact points. These tools are based on the definition of a *Grasp Matrix* $\boldsymbol{G}$ and a *Hand Jacobian* $\boldsymbol{J}$. Let n_c be the number of contact points and n_q be the number of joints in the hand, the Grasp Matrix $\boldsymbol{G}$ maps the object twists to the transmitted twists in

Journal Name	Foundation year	H-index	uncited publications after 1 year
IEEE Transactions on Robotics (T-RO)	2004	121	9.5%
International Journal of Robotics Research (IJRR)	1982	128	13.5%
Autonomous Robots	1994	91	19.46%
Journal of Field Robotics (JFR)	2006	77	19.7%

Table 3: Top journals in Scientific Robotics

each contact point, while the Hand Jacobian $\boldsymbol{J}$ maps the joint velocities to the transmitted contact twists on the hand. Let $\boldsymbol{\nu}$ be the twist of the object with respect to a global reference, $\boldsymbol{\nu_{c,obj}}$ be the transmitted twists on the object expressed with respect to each contact point, $\boldsymbol{\nu_{c,hnd}}$ the same on the hand and $\dot{\boldsymbol{q}}$ the joint velocities, then we have:

$$\boldsymbol{\nu_{c,obj}} = \boldsymbol{G}^T\boldsymbol{\nu}$$
$$\boldsymbol{\nu_{c,hnd}} = \boldsymbol{J}\dot{\boldsymbol{q}}$$

The choice of what twists are transmitted between object and hand encodes the contact model, most used are **Point-contact-without-Friction** (only normal component transmitted, and no momentum), **Hard Finger** (all translational components, no momentum) and **Soft Finger** (all translational components and normal momentum). These matrices only depend on the contact point geometry and hand configuration, and encode all the information about the grasp. The quantification of the actual closure and robustness of a grasp configuration can be encoded into a linear programming problem, and thus efficiently extracted. We refer to [14] for an in depth analysis on this topic.

Many optimization algorithms and simulation tools have been devised around this, in particular we refer to *GraspIt!* which collects a number of evaluation and optimization tools in an open simulated environment [11]. It is nonetheless very popular to find data-driven approaches in research on grasping. No specific machine learning model dominates the scene, but still very popular tools for grasping and affordance learning are deep learning, reinforcement learning and bayesian approaches in general.

1.2. Research topic

The Robotic Grasping research field tackles the problem of automating grasping actions on novel objects under different sensorimotor conditions. It follows the classical framing of sense, plan and act, thus the perception (intended as sensing and modeling the environment) is the basis on which the following steps must base. The challenges and opportunities of our research can be better framed in the more general context of affordance perception. Since the first definition of affordances, by Gibson in 1966, [7] a long discussion evolved, for a complete discussion refer to [18].

According to Michaels [10], affordances are emergent properties embodied in the relations between an animal and its environment directly connected with the possibility of action of the animal with the environment. Applied to robotics, affordance perception means understanding the possibility of action of a robot depending on the possible relations between its actuators and the environment to achieve high level tasks. In this context grasping represents an affordance for the control of some or all degrees of freedom of some object with a hand-like physical actuator.

The possibility of having control on some degrees of freedom of objects is fundamental for robotics applications as it is usually the main goal of actions. Moreover, in the wider context of affordances, task-oriented grasping enables the possibility of tool use, which in turns allows an enormous range of new possibilities of action. Being able to model and understand the environment is a critical point to plan a solid grasp, and even more to relate the grasp with a task; for this reason research in Perception for Robotic Grasping is necessary to achieve general task-oriented grasping.

2. Main related works

2.1. Classification of main related works

The research community in Perception for Robotic Grasping has proposed very different and heterogeneous perceptual models, within very different settings. Focusing only on the perception and modeling of the environment we can notice a strong correlation between the perceptual basis of a model and its limitations. In particular we classify the main related works according to the *dimensionality* of the vision system which models the environment:

- **Blind Perception**: in this category we include all systems which do not employ vision. It is important to underline that here we mean vision in its most general sense, as any perceptual system that can provide a global view of the environment, as a laser system could do. As such, blind systems have only access to local information about the contact points of the fingers, with the consequent strong limitation of not being able to plan any new grasp. The purpose of these works is usually to evaluate or improve an existing grasp to make it more solid via contact information like local pressure maps or motors torque.

- **Appearance Perception**: in this category we include all systems employing at least a monocular vision system or equivalent. They do not necessarily perceive nor model any clue of absolute spacial dimensions and shapes on which to plan grasps, although they have some global perception of the environment.

- **Geometry Perception**: there are many ways a system can employ to estimate real distances and have a notion of three-dimensional space and object shape, ranging from RGB-D cameras to model priors. All works that perceive or model information connected with metric distances about the environment have a clear extra opportunity as they can employ geometrical models to plan accurate grasps. Inside this broad category we can further differentiate according to the span of the spacial model:
  - **Focused**: these works try to estimate or consider a detailed spacial model of the object to be grasped
  - **Global**: these works try to model the whole environment and perceive many objects at once, usually but not necessarily with greater uncertainty than focused models.
  - **Mixed**: these works estimate a model of the environment and also identify distinct object models and their locations

- **Physics Perception**: a further dimension of perception of the environment for its understanding includes the physical perception. Its relevance in understanding the environment comes from the predictive power of physics as a model of the interaction with objects. As a consequence, physical quantities efficiently encode the knowledge of the agent about the evolution of the environment in time and as such become a general perceptive model of the "*dynamic state*" of objects, as shape and position encodes their spacial "*static state*". In this category we deliberately leave apart the perception of local physical entities like touch pressure as they do not provide any global information about the environment and alone should be regarded as Blind Perception. The reason for this is the limited exploitability of such information, as already discussed in the appropriate section.

We must notice that this classification is partially hierarchical: a model exploiting physics perception must perceive also geometry, and in turn a model perceiving geometry also perceives appearance. The category of Blind Perception instead includes all systems that do not access global information and thus cannot be included in the others. When classifying a work we will label it with the most restrictive category in which it can be included.

2.1.1 Blind Perception

Blind perception research works under the assumption of knowing only local contact or joint information, with no concern about the environment. Some works like Arimoto et al. [1, 2] deviate from the usual objective of statically stable closures and suggest a control theory approach on two-parallel-finger grippers, which cannot achieve force closure statically. Although interesting, the idea is limited by the need of grippers with fine and strong finger control and only two parallel fingers, which is impractical. As a result, the authors could validate their methods theoretically and by simulations, but no experiment on real robots has been done.

More recent works like Dang et al. [5, 4] pose the objective of evaluating or improving a given grasp in terms of its closure and robustness. In their first work [5] the authors provide a machine learning approach using Support Vector Machines to evaluate the robustness of a grasp based on tactile feedback. Data are collected through simulation, and the generalization of their approach has been tested again only in a simulated environment with a model of the three-fingered Barrett Hand [17]. Their next work [4], instead, subsumes the task of evaluation by trying to *improve* a given grasp. Dang uses the same pattern of collecting significant data from simulations and exctracting operative knowledge from them, this time through the use of a K-nearest neighbors. By mimicing a hand pose and joint configuration similar to the known stable grasps which are nearest to the current grasp, the hand is supposed to end to a more robust configuration. The authors proved, with experiments on a real Barrett hand, to randomly grasp a novel object, and they showed this blind policy can lead to significant improvements in the robustness of grasps.

A common limitation of all these works, which is intrinsic to the category, is the inability to plan any new grasps on objects whose position in space is not known a priori. A good opportunity, instead, comes from the reduced (though not insignificant) uncertainty of local measurements, which allows good generalization of simulation data to real applications.

2.1.2 Appearance Perception

The works under this category consider global knowledge of the world, but no explicit geometrical notion of distance and space is considered. This category received very limited attention from the research community and only few works are available as the great majority of researchers considering vision explicitly model and estimate at least some key spacial cues.

A relevant work in this context is the one by Levine et al. [9] who trained a deep learning based controller by leveraging on an extremely large scale data collection phase. The robots they used are arms with two-finger grippers positioned in front of a box with different small objects, with a single RGB uncalibrated camera facing towards the box. The learned controller had the objective of successfully gripping every time a different object to lift it and place it back again. The authors proved that a learned controller could generalize well on different camera illumination and calibration conditions and different finger tension or tearing levels.

This work shows that under constrained settings it is possible to extract an implicit model of the required information for grasping from data even if the input information is extremely uncertain, variable, and incomplete. We must still take into consideration that this approach is limited by the extreme effort in collecting the required data and by the inability to produce a single model that generalizes over different tasks and settings.

2.1.3 Geomtry Perception

Most of the works in Robotic Grasping belong to this class, thus very different approaches and settings have been defined. Some models extract just essential spacial information, like Kim et al. [8] who roughly estimate the 3D position of the target object from stereo vision for the first gross hand approaching movement. They use the interesting idea of positioning a stereo camera on the hand itself, to be able to detect the target in the environment and then focus on it while approaching, thus we can classify this work as being *mixed span*. Another important contribution is the one from the MIT Princeton team at the Amazon Challenge 2017 [19]: their work has analogies with the one from Levine [9] as it works in unstructured environments with many objects cluttered in a container, but they use four RGB-D cameras for a more informative perception with global span and directly infer gripper or suction affordance maps from deep models while employing simpler arm controllers to move: in this case space, differently from Levine, is explicitly known.

Other works in this category go in the direction of grasping complex objects after the analysis of their surface. Erkan et al. [6] propose to detect short segment edges on the surface of the object through Early Cognitive Vision descriptors and classify pairs of coplanar segments, according to the quality of the grasp they afford, through semi-supervised learning. Their approach has a clearly focused span on a single object whose interesting surface features are mapped in the space and jointly suggest grasp possibilities.

The works seen so far do successfully grasp objects, but they ignore completely that different grasps are required for different tasks: they fix the task of grabbing and eventually moving a target object, but they can hardly be generalized to different tasks. Biasing grasps towards the completion of some task is a key aspect for the relevance of grasping as discussed in Section 1.2. One of the first works integrating grasp planning with tasks is Prats et al. [13]. They use simple loaded 3D models of home objects (doors, drawers, windows) and preshapes of standard hand configurations to enact some task encoded in physical interactions like applying force or torque on specific degrees of freedom of the object. Their heuristic method has been successful on experiments with a real Barrett hand with a very rough model of the target object, but still it requires a 3D model of the object and it does not account for the high uncertainty of directly perceiving the model. We do not classify this work as Blind Perception as it takes into account geometrical knowledge of the environment, even if endowed and not perceived, nor it falls in the category of Physical Perception as physical quantities are used to model the task, not the object, although through preshapes they are the primary link between a task and its associated grasp.

2.1.4 Physics Perception

The physical understanding of the environment is recently receiving attention from the Computer Vision community, trying to estimate various quantities such as mass [16], material [15, 3] or manipulation forces [12]. However, rarely researchers in the Robotic Grasping fields use similar techniques to physically model the environment.

A significant work in our analysis is the one by Zhu et al. [20], in the more general framework of affordance learning, but still relevant for task-oriented grasping. They provide a model to estimate the best suited tool for a task among the ones presented on a planar surface through vision. The task is presented to the system through a video of a human demostrator choosing the best tool among a different set and using it to perform a task like nutcracking. The model involves the optimization of physical quantities in imagined tool uses, and the choice of the best area of the tool for grasping and for functional use. The method is validated through benchmarking of the system choices compared against the ones taken by humans. The main limitation in this is the lack of an intrinsic definition of task which would enable further elaboration and adaptation to robotic actuators which are different from humans; moreover we miss a concrete connection between the planned grasp area on the tool and the actual grasp pose to effectively execute the task.

2.2. Conclusions

As we have seen, the problems of evaluating the robustness of a grasp and planning grasps with perfectly known object models have been fully assessed, while consistent research efforts are currently producing good results towards the same problems under uncertain object models. On the contrary, the problem of task-oriented manipulation is still an open problem in the Robotic Grasping field: few works have been attempted, lacking a generalized framing of tasks either limiting task expressivity [20] or categorizing action possibilities [13]. Moreover, the modeling of physical quantities, which explain the connection between actions, tools and tasks, received almost no attention from the Robotic Grasping community and remains an unexplored opportunity.

References

[1] Arimoto, S., Ozawa, R., and Yoshida, M. Two-dimensional stable blind grasping under the gravity effect. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (April 2005), pp. 1196–1202.

[2] Arimoto, S., Yoshida, M., and Bae, J.-H. Stable "blind grasping" of a 3-d object under non-holonomic constraints. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* (May 2006), pp. 2124–2130.

[3] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3479–3487.

[4] Dang, H., and Allen, P. K. Tactile experience-based robotic grasping. In *Workshop on Advances in Tactile Sensing and Touch based Human-Robot Interaction, HRI* (2012).

[5] Dang, H., Weisz, J., and Allen, P. K. Blind grasping: Stable robotic grasping using tactile feedback and hand kinematics. In *2011 IEEE International Conference on Robotics and Automation* (May 2011), pp. 5917–5922.

[6] Erkan, A. N., Kroemer, O., Detry, R., Altun, Y., Piater, J., and Peters, J. Learning probabilistic discriminative models of grasp affordances under limited supervision. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (2010), IEEE, pp. 1586–1591.

[7] Gibson, J. J. The senses considered as perceptual systems.

[8] Kim, D.-J., Lovelett, R., and Behal, A. Eye-in-hand stereo visual servoing of an assistive robot arm in unstructured environments. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on* (2009), IEEE, pp. 2326–2331.

[9] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research 37*, 4-5 (2018), 421–436.

[10] Michaels, C. Affordances: Four points of debate. *ECOLOGICAL PSYCHOLOGY 15* (04 2003), 135–148.

[11] Miller, A. T., and Allen, P. K. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics Automation Magazine 11*, 4 (Dec 2004), 110–122.

[12] Pham, T.-H., Kheddar, A., Qammaz, A., and Argyros, A. A. Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 2810–2819.

[13] Prats, M., Sanz, P. J., and Del Pobil, A. P. Task-oriented grasping using hand preshapes and task frames. In *Robotics and Automation, 2007 IEEE International Conference on* (2007), IEEE, pp. 1794–1799.

[14] Prattichizzo, D., and Trinkle, J. Grasping. In *Handbook on Robotics*, B. Siciliano and O. Kathib, Eds. Springer, 2008, pp. 671–700.

[15] Sharan, L., Liu, C., Rosenholtz, R., and Adelson, E. H. Recognizing materials using perceptually inspired features. *International journal of computer vision 103*, 3 (2013), 348–371.

[16] Standley, T., Sener, O., Chen, D., and Savarese, S. image2mass: Estimating the mass of an object from its image. In *Conference on Robot Learning* (2017), pp. 324–333.

[17] Townsend, W. Mcb—industrial robot feature article—barrett hand grasper. *Industrial Robot: An International Journal 27*, 3 (2000), 181–188.

[18] Zech, P., Haller, S., Lakani, S. R., Ridge, B., Ugur, E., and Piater, J. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior 25*, 5 (2017), 235–271.

[19] Zeng, A., Song, S., Yu, K.-T., Donlon, E., Hogan, F. R., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E., et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), IEEE, pp. 1–8.

[20] Zhu, Y., Zhao, Y., and Chun Zhu, S. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 2855–2864.