Simulation Empowered Self-Supervised Deep Learning for Adaptive Robotic Grasping

Mastering robotic grasping is a necessary skill for a robot to perform tasks involving the manipulation of one or more objects. As automation increases, these tasks are nowadays found in more and more fields of industry such as car manufacturing, waste sorting, or food processing. In such cases, the environment cannot be fully controlled, and it is, therefore necessary to use systems capable of analyzing their environment and interacting with them, instead of just doing a set of preprogrammed tasks.

To do this, detecting the configuration of instances of objects in a scene using their 3D models is not always possible. In some cases, the models are not available. It is then possible to use a geometric approach to detect opportunities for grasp positions in images. However, these approaches require many parameters to be set manually so that the geometric criteria are adapted to the scene. To overcome this problem, methods based on deep learning can be used: from a large number of annotated images and good grasping positions, a deep neural network can learn how to predict good positions on new images. The acquisition of such a large amount of annotated data, however, represents an obstacle for robotics.

In collaboration with Siléane, a French industrial robotics company, the aim of this work is therefore to develop solutions for learning robotic grasping through simulated data, which are available in large quantities. This work has two main contributions.

Firstly, we propose a new neural network architecture that predicts grasp positions for a parallel-jaw gripper on images of isolated objects. Previous state-of-the-art work used small amounts of manually annotated data. In this work, we build a new large-scale dataset of synthetic images with automatically generated annotations through physics simulation and train our network using it. The use of a large amount of diverse data, rather than just a few images, allows the network to be trained on a wider range of situations, and thus be able to handle more different unknown cases.

Secondly, the work presented here deals with the detection of grasp locations in bin-picking context, *i.e.* in scenes with many objects occluding each other. While traditional approaches for this problem use local information, classifying the potential quality of a grasp according to the surrounding data in the image, our proposed network completes this information by adding the notion of object instances. Trained on self-supervised simulated images, it can thus estimate the quality of a grasp position based not only on local information, but also on the global context of the object present at the considered position in the image. Based on extensive experiments, we show that this double approach allows improving the quality of predictions, both in a simulated environment and in real robotic tasks.

**Keywords**: deep learning, simulation, robotic grasping, bin-picking, neural networks, synthetic data

Apprentissage profond auto-supervisé par simulation pour la saisie robotique adaptative

La maîtrise de la préhension robotique par un robot est nécessaire pour l'accomplissement de toutes les tâches nécessitant la manipulation d'un ou plusieurs objets. Avec une automatisation croissante de l'industrie, ces tâches se retrouvent aujourd'hui dans de nombreux domaines de l'industrie tels que l'automobile, le tri des déchets ou encore l'agro-alimentaire. Dans de tels cas, l'environnement ne peut pas être totalement contrôlé, et il est donc nécessaire de faire appel à des systèmes capables d'analyser leur environnement pour interagir avec eux.

Pour ce faire, on ne peut pas toujours utiliser les modèles 3D des objets pour détecter la configuration des instances dans une scène. Dans certains cas, les modèles ne sont pas disponibles. Il est alors possible d'utiliser une approche géométrique pour détecter des opportunités de prises robotiques dans des images. Cependant, ces approches nécessitent de régler de nombreux paramètres manuellement pour que les critères géométriques soient adaptés à la scène. Pour pallier à ce problème, il est possible d'utiliser des méthodes à base d'apprentissage automatique : à partir d'un grand nombre d'exemples d'images et de bonnes positions de prises, un réseau de neurones profond est capable d'apprendre à prédire des bonnes positions sur de nouvelles images. L'acquisition d'une telle quantité de données annotées représente cependant un obstacle pour la robotique.

En collaboration avec Siléane, une entreprise de robotique industrielle française, l'objectif de ce travail est donc de développer des solutions pour l'apprentissage de la préhension robotique à travers les données simulées, disponibles en grandes quantités. Dans ce domaine, ce travail apporte deux contributions.

Premièrement, nous proposons une nouvelle architecture de réseau de neurones permettant de prédire des positions de prises pour une pince à mors parallèles sur des images d'objets isolés. Les précédents travaux de l'état de l'art utilisaient de faibles quantités de données annotées manuellement. Dans ce travail, nous construisons une très grande base de données d’images synthétiques annotées automatiquement par simulation physique, que nous utilisons ensuite pour entrainer notre réseau. L'utilisation d'une grande quantité de données diversifiées, plutôt que de quelques images seulement, permet au réseau d'être entrainé sur des situations plus variées, et ainsi de pouvoir gérer de plus nombreux cas différents.

Dans un deuxième temps, les travaux présentés ici s'intéressent à la détection de prises au sein d'un vrac, *i.e.* d'un enchevêtrement de nombreux objets avec de forts recouvrements entre eux. Alors que les approches traditionnelles dans ce domaine utilisent une vision locale, classifiant la qualité potentielle d'une prise en fonction des données alentours dans l'image, le réseau proposé complète cette information en ajoutant la notion d'instances d'objets. Entrainé sur des images simulées de manière auto-supervisée, il peut ainsi estimer la qualité d'une position de prise en se basant non seulement sur une information locale, mais également sur le contexte global de l'objet présent à la position considérée dans l'image. A partir de plusieurs expériences, nous montrons que cette double approche permet d'améliorer la qualité des prédictions, aussi bien dans un environnement simulé que dans de vrais contextes robotiques.

**Mots-clés :** apprentissage profond, simulation, préhension robotique, dévracage, réseau de neurones, données synthétiques, auto-supervision