Multiple Pregrasping Poses Prediction Using Combining Deep Convolutional Neural Network and Mixture Density Network

Sungphill Moon, Youngbin Park, and Il Hong Suh

Department of Electronics and Computer Engineering, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, Korea
sp9103@incrol.hanyang.ac.kr,pa9301@hanyang.ac.kr,ihsuh@hanyang.ac.kr

Abstract. In this paper, we propose a deep neural network to predict pregrasp poses of a 3D object. Specifically, a single RGB-D image is used to determine multiple 3D positions of three fingers which can provide suitable pregrasps for a known or an unknown object in various poses. Multiple pregrasping pose prediction is typically complex multi-valued functions where standard regression models fail. To this end, we proposed a deep neural network that contains a variant of traditional deep convolutional neural network, followed by a mixture density network. Additionally, to overcome the difficulty in learning with insufficient data for the first part of the proposed network we develop a supervised learning technique to pretrain the variant of convolutional neural network.
abstract environment.

Keywords: Grasping pose prediction, deep convolutional neural network, mixture density network

1 Introduction

In the review of Sahbani et al. [1], grasping methods can be broadly categorized as analytic and data-driven Most analytic methods assume the availability of complete knowledge of the objects to be grasped, such as the complete 3D model of the given object. These methods then construct a suitable grasp pose based on criteria, such as force closure or stability. Therefore, grasp synthesis is usually formulated as a constrained optimization problem over those criteria. Therefore, analytic methods typically need to understand the precise 3D shape of the object and require huge computation for solving the optimization problem.

Data-driven approaches, on the other hand, investigate the way to avoid such disadvantages by imitating human grasping. These methods select a appropriate grasp by means of building a direct mapping from vision to action. A majority of these methods have hence more focus on use of vision-based features obtained from RGB or RGB-D images to predict grasp locations. Learning visual feature based on machine learning algorithm have enabled grasp pose estimation to generalize easily to novel objects encountered often in uncontrolled environments.

The methods that capture the mapping from vision to action by a deep learning model [3], [4] has recently gained much attention thank to recent immense success of deep learning in a wide variety of tasks, including robotic grasping and manipulation [6], [5] as well as object recognition [7], semantic segmentation [8], caption generation [9]. However, the main difficulty for deep learning is that training deep neural network requires a large-scale data collection.

Fig. 1. First three images on the left side and last three images on the right side show two qualitatively different pregrasp poses. In first three images, thumb appears in the upper part of the image while in last three images thumb appears in the lower part of the image. Three pregrasp poses on the left and right sides seem to be identical within own group but they are slightly different.

Figure 1 shows pregrasp poses. We refer a pregrasp pose as the configuration where closing the fingers until resistance is encountered can leads a proper grasp pose. In this paper, we address the problem of multiple pregrasp poses regression of a 3D object using deep neural network. Specifically, a single RGB-D image is used to determine multiple 3D positions of three fingers which can provide suitable pregrasps for a known or an unknown object in various poses. To this end, we first create a considerably large number of human annotated pregrasp data. In the dataset, an image containing only a specific pose of an object is used as an input and the corresponding 150∼300 human-supervised pregrasp poses are used for a set of labels. Here, at least two qualitatively different finger configurations are included in the a set of labels.

In this case, it is extremely hard to collect a great amount of training data using traditional kinesthetic teaching procedure, where the human teacher directly moves the robotic arm to make the robot performs pregrasp. To overcome this problem, we detached robotic hand from the robot arm and attached optical markers to ends of three fingers to track 3D positions of the fingers using optical motion capture system. The human teacher then grabbed the robotic hand and demonstrated possible pregrasp poses for a while, varying poses with small continuous movements. At this time, various 3D positions of three fingers are recoded via motion capture system. Again, a qualitatively different finger configuration for the object pose is provided by the human teacher and identical demonstration procedure repeated. Figure 1 shows our data acquisition procedure. It should be considered that all images in Figure 1 are not subject for training data. Only 3D positions of three figures are recorded for targets. Optical markers are omitted in the figure for displaying robotic hand clearly.

The advantage of this data collection procedure is that much less efforts are necessary to gather same amount of training data compared to traditional kinesthetic teaching mentioned before. In this paper, we investigate pregrasp poses prediction instead of the estimation of grasp poses to exploit this data collection scenario. The rationale behind this is that the accurate pregrasp leads successful robotic grasping with high probability.

We build our model primally based on traditional deep convolutional neural network (DCNN). However, DCNN only is not sufficient to model robotic pregrasp especially in case that training dataset contains the data where there are multiple possible pregrasps for a specific pose of an object. To address this inherent limitations, we investigate a model that combines DCNN and mixture density network (MDN) [10]. MDN has been shown to be successful for complex multivalued functions where standard regression models fail. The combined DCNN and MDN is trained using the human annotated pregrasp dataset gathered by proposed data collection procedure. It should be noted that in our dataset there are a large number of pregrasp pose labels but relatively small number of images containing objects are provided. In this case, DCNN cannot learn rich visual feature to predict suitable pregrasp location. Therefore, we initialize DCNN using the proposed supervised pretraining method.

2 Related Works

In this section, we review several robotic grasping literatures on data-driven approaches, in which we focus more on the studies that investigate use of deep learning for prediction of grasp location. For a comprehensive survey of robotic grasping, we refer the reader to recent surveys on the subject [1], [2].

The robot's own trial and error experiences is a one way to collect training data for grasp tasks [11], [12]. However, performing more than a few hundred trial and error runs by a physical robot is usually difficult. Subsequently, such small dataset often causes the machine learning models to overfitting. On the other hand, human annotated benchmark dataset are used to train deep neural network in a supervised way [13]. Cornell grasping dataset used in the study contains 1035 images of 280 graspable objects. Each image is annotated with several ground-truth positive and negative grasping rectangles. Each of these patches is fed to the network to evaluate grasp quality scores. Therefore, several feed-forward computations are performed to determine best grasp pose.

Self-supervised learning of grasp poses is another way to train deep neural network in a supervised way. Pinto and Gupta proposed a self-supervised data collection method without human supervision using a heuristic grasping system based on object proposals [3]. They used most recently learned model to gather data so that data collection procedure become more and more efficient. The dataset has more than 50K datapoints and has been collected using 700 hours of trial and error experiments using the Baxter robot having two grippers. After training, given an image patch, an 18-dimensional likelihood vector where

each dimension represents the likelihood of whether the center of the patch is graspable at 0°, 10°, . . . 170° is estimated.

Levine et al. [4] present a self-supervised learning method similar to the work proposed by Pinto and Gupta. However, the training dataset consists of over 800,000 grasp attempts on a very large variety of objects, which is more than an order of magnitude larger than the dataset collected by Pinto and Gupta. To obtain this huge dataset, it has taken two months, using between 6 and 14 robotic manipulators at any given time. A deep convolutional neural network called grasp success predictor has been trained to determine how likely a given motion is to produce a successful grasp. The performance of this method is state-of-the-art for robotic grasping, but data collection procedure is time-consuming and as in the work of [13], several feed-forward computations are required to determine next movement.

Most closely related deep neural network structure to our proposed neural architecture is proposed by Levine et al. [5] In this work, a novel DCNN architecture is investigated. Specifically, the first half of the network contains three convolutional layers, followed by a spatial softmax and an expected position layer that converts pixel-wise features to feature points. The expected position layer provides accurate spatial reasoning and reduces the number of parameters to avoid overfitting. The rest half of the network is developed to generate motor torques of the robot arm for various object manipulation tasks. In this paper, we adapt the idea converting pixel-wise features to feature points to build our proposed deep neural network.

3 The Proposed Neural Network Architecture

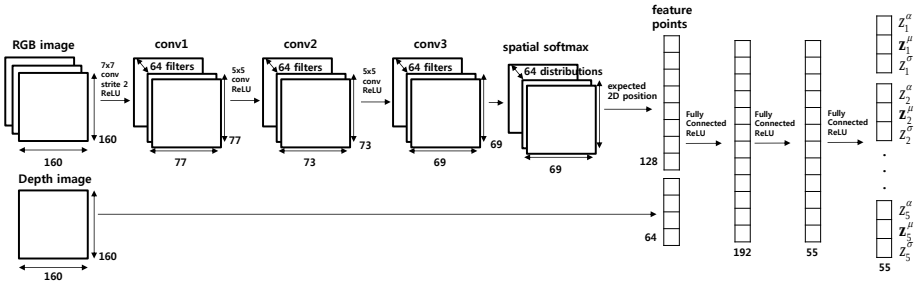

Fig. 2. The Proposed Neural Network Architectures.

The proposed neural network architecture is illustrated in Figure 2. The input of the network is a 160 x 160 RGB image that contains only an object. First part of the network contains three convolutional layers, followed by a spatial softmax and an feature points layer that converts pixel-wise features to expected positions. The spatial softmax layer provides lateral inhibition, which suppresses low

activations and keeps only strong activations that are more likely to be accurate. The feature points layer computes expected position (x, y) of each channel in softmax layer. As shown in Figure 2, depth of each expected position is added in feature points layer. The dimension of feature points layer is thus three times more than the number of channel in softmax layer. In particular, adding depth information does not play an important role and leads only slight improvement in performance. As mentioned before, this part of network except adding depth data is analogous to the network architecture proposed by Levine et al [5]. In the following, we will refer the first part of the neural network to DCNN$^+$ for simplicity. We found that the proposed architecture achieves better performance on the prediction of pregrasp poses compared to the architecture where DCNN$^+$ is replaced by traditional DCNN. The spatial softmax and an expected position layers are closely described in [14]. Second part of the network is mixture density network (MDN) and is composed of three fully connected layers. In the following sebsections, we will present the details of our two contributions: MDN and pretraining of DCNN$^+$.

3.1 Mixture Density network

An MDN combines a mixture model with an artificial neural network. This paper employs Gaussian mixture (GMM)-based MDN to predict multiple pregrasp poses. As shown in Figure 2, a set of 3D feature points is input for the MDN used in the proposed neural network architecture. The activations in output layer is in turn transformed to the parameters of a GMM. The GMM parameters can be derived from the MDN as

$$\alpha_i = \frac{exp(z_i^\alpha)}{\sum_{j=1}^{m} exp(z_j^\alpha)}, \quad \mu_{ik} = z_{ik}^\mu, \quad \sigma_i = exp(z_i^\sigma). \tag{1}$$

Here, the parameters for the i-th Gaussian, mixture weight, mean and variance are denoted by α_i, μ_i and σ_i, respectively. m denotes the number of kernel in GMM and in our experiments we used 5 Gaussians. In case of multivariate GMM, $\boldsymbol{\mu}_i$ is the vector, with component μ_{ik}. $\boldsymbol{\mu}_i$ in this work is 9 dimensional vector because the pregrasp pose in this work is represented by 3D positions of three fingers. Although $\mathbf{z}_i^\mu$ shown in Figure 2 is illustrated as scalar for visualisation purpose but it is 9 dimensional vector as well. The covariance matrix for the i-th Gaussian denoted by $\boldsymbol{\Sigma}_i$ and is equal to $\mathbf{I}\sigma_i$. The rationale behind this is that the components of $\boldsymbol{\mu}_i$ can be assumed to be statistically independent, and can be described a common variance σ_i. The details of this discussion are described in [10]. Full probability density function of an pregrasp pose $\mathbf{t}$, conditioned on a set of 3D feature points $\mathbf{x}$ is given as

$$p(\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x})\phi_i(\mathbf{t}|\mathbf{x}) \tag{2}$$

where

$$\phi_i(\mathbf{t}|\mathbf{x}) = \frac{1}{(2\pi)^{c/2}\sigma_i(\mathbf{x})^c} exp\{-\frac{\|\mathbf{t}-\boldsymbol{\mu}_i(\mathbf{x})\|^2}{2\sigma_i(\mathbf{x})^2}\}. \quad (3)$$

Here, c represents the dimension of input vector. The loss of training of the MDN is to minimize the negative log likelihood given the training data as

$$\ell = \sum_{j=1}^{n}\left[-ln\{\sum_{i=1}^{m}\alpha_i(\mathbf{x}^j)\phi_i(\mathbf{t}^j|\mathbf{x}^j)\}\right] \quad (4)$$

where $(\mathbf{x}^j, \mathbf{t}^j)$ is a j-th input/target pair and n is the number of training data.

3.2 Pretraining

Our training dataset collected using the method described in Introduction has more than 100,000 pregrasp poses but there are only 550 images that contain object poses. To learn rich and valuable visual features can predict the suitable pregrasp pose, we thus need to pretrain DCNN$^+$ using more training data that include various objects and a large number of different object poses. Finn et. al [14] proposed unsupervised learning algorithm to pretrain identical network structure, in which loss function is to minimize reconstruction error. However, we found that this unsupervised learning often produces more than half of the feature points on the background. To overcome this limitation, we proposed supervised pretraining algorithm. The output of the network for pretraining is the center position (cx, cy) of the object in the image. We collected additional dataset for pretraining. The center position of the object was estimated by using simple background subtraction algorithm. For supervised training, 19 objects were used and 54,000 images were captured. To collect such amount of dataset easily, we threaded objects and pulled the objects in several different directions. At that time, the RGB-D camera continued to record the scenes.

4 Experiment

4.1 Experimental Setup

Dataset consists of 8 categories of objects such as *cup*, *cellphone*, *pen*, *doll*, *lotion*, *can*, *small cylinder*, *toy block*. Three different objects were included for each category in training dataset. The three object used in training and two more objects were included for each category in test dataset. Figure 3 shows all objects used for our training and test data. The objects in the right table are training data and the objects in the left table are objects that are not shown in training phase. The objects in the right table were employed for test as well but they are shown in different poses to the ones in the training phase. The size of workspace is approximately 1m x 1m. In training phase, an object was placed in 5 different positions and was rotated 3∼8 times at the each position. 150∼300 human-supervised pregrasp poses were labeled for an input. Therefore, the number of

Fig. 3. All objects used for our training and test data.

input images where each of them includes only an object is 550 and the number of target pregrasp poses for the inputs is 119.243. For test dataset, a known object appeared 6∼7 different poses and a unknown object was placed in 5 different poses. There are hence 20 and 10 test data for each known and unknown object, respectively. The total numbers of test data for known and unknown objects are 160 and 80, in that order. 100∼150 pregrasp poses are provided for ground truth of an input in test dataset. For an input image, two qualitatively different pregrasp poses are demonstrated by a human teacher in both training and test dataset.

We have implemented three methods for pregrasp pose regression for the sake of comparison: (1) DCNN$^+$+SP (Pretraining DCNN$^+$ using the proposed supervised learning method), (2) DCNN$^+$+USP+MDN (Pretraining DCNN$^+$ using the unsupervised learning method proposed in [14] and combining DCNN$^+$ and MDN) and (3) DCNN$^+$+SP+MDN (Pretraining DCNN$^+$ using the proposed supervised learning method and combining DCNN$^+$ and MDN). DCNN$^+$+SP+MDN is the full implementation of our proposed methods.

We implemented the three methods based on Caffe. The number of epochs for three methods in training phase are 35,000 and the training took approximately 18 hours on a system equipped with a NVIDIA TITAN GPU. All experiments were conducted using the robotic hand with three fingers and the Microsoft Kinect v2 was exploited for RGB-D camera.

4.2 Quantitative Results

In this sebsection, we present the numerical evaluation and comparison of the performances of the three methods. We measured the performance on pregrasp pose prediction using average error (AVE). While DCNN$^+$+SP predicts only one pregrasp pose two methods that combine DCNN$^+$ and MDN produce multiple pregrasp poses. As illustrated in Section 3, the maximum number of the pregrasp poses produced by MDN is 5. We can consider the i-th pregrasp pose reliable if the corresponding α_i in Equation 1 is larger than a certain threshold γ. γ was set to 0.3 in this experiment. Then, for computing AVE, we should select one pregrasp pose among the reliable predictions.

As mentioned before, a test data has 100∼150 true pregrasp poses. For computing AVE, we also select one ground truth which has the minimum euclidean distances relative to the prediction. Euclidean distance for each finger is then computed between the prediction and the selected ground truth. Average of the three euclidean distance is determined as the error for the test data.

Table 1. Average pregrasp pose prediction error. DCNN$^+$+USP+MDN and DCNN$^+$+SP+MDN select the mean of the largest Gaussin as the precition

	DCNN$^+$+SP	DCNN$^+$+USP+MDN	DCNN$^+$+SP+MDN
AVE(known)	6.13cm	5.85cm	1.69cm
AVE(unknown)	6.52cm	5.79cm	2.53cm

Table 1 shows AVEs of three methods, where $\boldsymbol{\mu}_i$ of highest α_i is determined for the prediction of DCNN$^+$+USP+MDN and DCNN$^+$+SP+MDN, respectively. It is observed that the errors two comparison methods are significantly increased compared to the proposed method. This result demonstrate effectiveness of the two proposed methods: the use of MDN and the supervised pretraining for DCNN$^+$.

Table 2. Average pregrasp pose prediction error. DCNN$^+$+SP+MDN select the mean of the second largest Gaussin as the precition

AVE(known)	1.8cm	AVE(unknown)	2.65cm

The table 2 illustrate the advantage of combining DCNN$^+$ and MDN. In this experiment, $\boldsymbol{\mu}_i$ of second highest α_i is determined for the prediction of DCNN$^+$+SP+MDN. It is noted that second reliable predictions produced by the networks achieve lower performances compared to the AVEs obtained by first reliable predictions but the decreases are just small. This experiment demonstrates the proposed network that combine DCNN$^+$ and MDN can produces suitable multiple pregrasp poses. In the case of DCNN$^+$+USP+MDN, the second highest α_i is never higher than 0.3 for all test data. This results support the hypotheses that if poor visual features are learned in DCNN$^+$ it is difficult to train MDN successfully.

4.3 Qualitative Results

Figure 4 illustrate the advantage of the proposed supervised pretraining. It is observed that most feature points obtained from the DCNN$^+$ which is pretrained using our proposed method are located on the object while more than

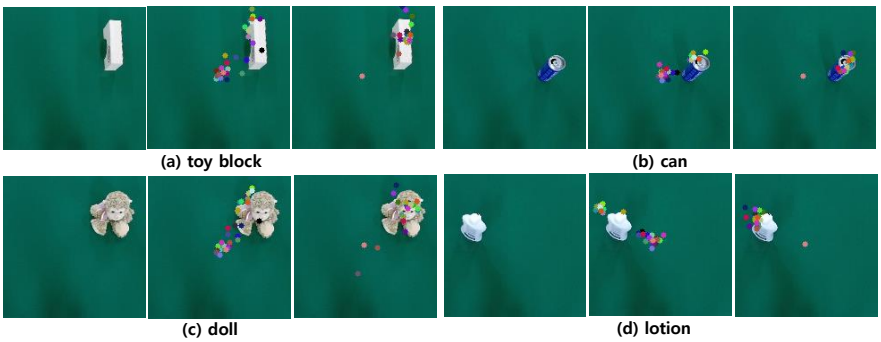

Fig. 4. Each left image of (a-b) is input image. Each center image of (a-d) displays values at feature points layer in Figure 2, in which DCNN$^+$ is pretrained using unsupervised learning. Each right image of (a-d) generated by pretraining DCNN$^+$ using the proposed supervised learning. Each feature is displayed in a different color.

half of the feature points generated from the DCNN$^+$ which is pretrained using unsupervised learning are located on the background. It seems that such small number of feature points come from objects can leads poor performance of DCNN$^+$+USP+MDN on the quantitative evaluations.

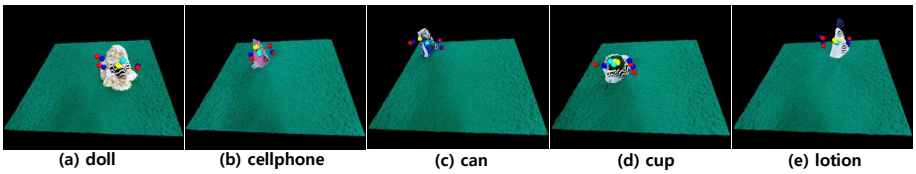

Fig. 5. (a-c) Known objects. (d-e) Unknown objects. Red dots represent ground truths, blue dots are the prediction produced by DCNN$^+$+SP+MDN, and yellow and sky blue dots are the predictions produced by DCNN$^+$+SP and DCNN$^+$+USP+MDN, respectively.

Figure 5 illustrate the advantage of the use of MDN. It is observed that 3D positions of three fingers predicted by DCNN$^+$+SP are located on the center of the objects while the grasp poses predicted by DCNN$^+$+SP+MDN are similar to the ground truths. DCNN$^+$+SP nearly predicts 3D positions of three fingers at the center of the object because in training phase such regression can minimize error between multiple grasp poses and a prediction.

5 Conclusions

In this paper, we presented a deep neural network architecture to predict multiple 3D positions of three fingers which can provide suitable pregrasps for a

known or an unknown object in various poses. To this end, we proposed a deep neural network that combines a variant of traditional deep convolutional neural network and a mixture density network. Additionally, to overcome the difficulty in learning with insufficient data for the variant of convolutional neural network we develop a supervised learning technique to pretrain the network.

We evaluated the performance of the proposed deep neural network against the two comparison methods. The results demonstrate the effectiveness of our method. Specifically, the use of MDN make possible to predict multiple pregrasp poses and the supervised learning pretraining enable rich and valuable visual features can predict the suitable pregrasp pose to be learned.

References

1. Sahbani, A., El-Khoury, S., and Bidaud, P. An overview of 3-D object grasp synthesis algorithms. Robot. Auton. Syst., vol. 60, no. 3, pp. 326-336, 2012.
2. Bohg, J., Morales, A., Asfour, T., and Kragic, D. Data-driven grasp synthesis-a survey. Robotics, IEEE Transactions on, vol. 30, no. 2, pp. 289-309, 2014.
3. Pinto, L. and Gupta, A. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. arXiv preprint arXiv:1509.06825, 2015.
4. Levine, S., Peter, P., Alex, K., and Deirdre, Q. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. arXiv preprint arXiv:1603.02199, 2016.
5. Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. JMLR., Vol. 17, no. 39, pp. 1-40, 2016.
6. Han, W., Levine, S., and Abbeel, P. Learning compound multi-step controllers under unknown dynamics. In Intelligent Robots and Systems (IROS), pp. 6435-6442, 2015.
7. Krizhevsky, A., Sutskever, I., and Hinton, G. Imagenet classification with deep convolutional neural networks. In NIPS, pp. 1097-1105, 2012.
8. Noh, H., Hong, S., and Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1520?1528, 2015.
9. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML, volume 37 of JMLR Proceedings, pp. 2048-2057, 2015
10. Bishop, C. Mixture density networks. Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University, 1994.
11. Detry, R., Baseski, E., Popovic, M., Touati, Y., Kruger, N., Kroemer, O., ... and Piater, J. Learning object-specific grasp affordance densities. in IEEE International Conference on Development and Learning (ICDL), pp. 1-7, 2009.
12. Paolini, R., Rodriguez, A., Srinivasa, S. S., and Mason, M. T. A data-driven statistical framework for post-grasp manipulation. The International Journal of Robotics Research, vol. 33, no. 4, pp. 600-615, 2014.
13. Lenz, I., Lee, H., and Saxena, A. Deep learning for detecting robotic grasps. The International Journal of Robotics Research, vol. 34, no. 4-5, pp. 705-724, 2015.
14. Finn, C., Tan, X. Y., Duan, Y., Darrell, T., Levine, S., and Abbeel, P. Deep Spatial Autoencoders for Visuomotor Learning. in IEEE International Conference on Robotics and Automation, 2016.