1 Motivations

1. Learning hand-eye coordination for robotic grasping from monocular images from scratch, with minimal prior knowledge and manual engineering.
2. End-to-end training directly from pixel input to output task-space gripper motion with minimal human supervision.
3. Precise camera calibration is not used.
4. The method learns servo the robotic gripper to position where grasps are likely to be successful.

2 Description of the approach

The approaches consists of two main components: a grasp success prediction network and a continuous servoing mechanism.

2.1 Grasp Success Prediction Network Network Description

1. The prediction network $g\left(\mathbf{I}_0, \mathbf{I}_t, \mathbf{v}_t\right)$ takes as input the current image $\mathbf{I}_t$, the image recorded before the grasp begins $\mathbf{I}_0$, and a command $\mathbf{v}_t$. The network outputs binary prediction if the task-space motion command will bring the gripper to a position where the fingers will pick up an object.
2. $\mathbf{I}_0$ provides view of the workspace without occlusion by the robot arm or gripper.
3. The command is a 5-element vector: a 3D translation vector in the coordinate frame of the robot's base, and a sine-cosine encoding of the change in the orientation of the gripper about the vertical axis.
4. Since the translation vector is in the frame of the robot's base, the approach does not require the precise calibration between the camera and the robot's end-effector.
5. The network is also an integrated and simpler approach, compared to previous modularized approach in grasping, where components to perform object localization and gripper localization are separated.
6. This is a complicated task since it requires the network to perform spatial reasoning to parse the geometry of the scene from monocular images, interpret the material property of objects and their spatial relationships.
7. The network also needs to infer the effects the task-space motion commands has on the position and orientation of the gripper.

2.2 Data Collection

1. Data for training the network is obtained by attempting to grasp using real physical robots.
2. Each grasp consists of T time steps.
3. At each time step, the robot records the current image $\mathbf{I}_t^i$ and the current pose $\mathbf{p}_t^i$ and choose a motion command for the gripper.
4. At the final time step T, the robot closes the gripper and evaluates the success of the grasp, producing a label ℓ_i.
5. Each grasp attempts result in T training samples, given by $\left(\mathbf{I}_t^i, \mathbf{p}_T^i - \mathbf{p}_t^i, \ell_i\right)$.
6. Each sample consists of the image observed at time t, the motion vector that brings the current pose to the final pose, and the success label.
7. The data collection process starts with random motor command selection and $T = 2$. The last command is always to close the gripper fingers. Random command was successful $10 - 30\%$ of the time.
8. T is gradually increased from 2 to 10 as the grasp prediction network is trained.
9. The objects were placed in front of the robot in metal bins with sloped slides to prevent the objects from becoming wedged into corners.
10. The objects are occasionally swapped out to increase the diversity of the training data.

2.3 Continuous Servoing

1. They use CEM to pick the motion vector that is likely to lead to successful grasp.
2. All samples from CEM are constrained to keep the final pose of the gripper within the workspace and prevent 180 degree rotation by the gripper.

2.4 Heuristics for lowering and raising the gripper

1. The gripper is raised if the grasp prediction network predicts that no motion has a probability of success that is less than 50% the probability of success of the best inferred motion.
2. The rationale is that if no motion is substantial worse than motion, then that means the gripper is stuck in an undesirable location and a large motion is required.
3. A large motion will hit neighboring objects.
4. Thus, raising the gripper will both get the gripper out of the undesirable location and prevent hitting neighboring objects.

2.5 Heuristics for deciding when to close the gripper fingers

1. They close the gripper finger when the grasp prediction networks predicts that no motion will have a probability of success that is at least 90% the probability of success of the best inferred motion.
2. The rationale is that the fingers should be close if no motion is as likely as moving to produce a successful grasp.

2.6 Heuristics for producing grasp success label

1. They use a combination of two mechanisms.
2. First, they check the state of the gripper fingers after each grasp attempt to see if the fingers close completely.
3. This test is effective at detecting large objects, but will miss small or thin objects.
4. To supplement this detector, they also use an image subtraction test.
5. They record an image of the scene after the grasp attempt (with the arm lifted above the workspace and out of view), and another image after attempting to drop the grasped object into the bin.
6. If no image is grasped, these two images are usually identical.
7. If an object was picked up by the gripper, then the two images will be different.

2.7 Assumptions

1. An important assumption is that the poses encountered in one grasp attempt has transitive relationship.
2. That is, moving from $\mathbf{P}_1$ to $\mathbf{P}_2$ and then to $\mathbf{P}_3$ is equivalent to moving from $\mathbf{P}_1$ to $\mathbf{P}_3$ directly.

3 Experiments

3.1 Baseline

1. They have two baselines: an open-loop baseline and a hand-engineered grasping system which uses depth images.
2. In the open-loop baseline, given an image of the scene, image patches are extracted. The image patch with the highest probability of success is chosen. A known camera calibration is then used to move the gripper to that location.
3. Comparison with this baseline is supposed to indicate the benefit of the continuous servoing mechanism.

4. The hand-engineered system uses a depth sensor instead of monocular camera. The point clouds obtained from the depth sensor were accumulated into a voxel map, which was then segmented using grasp-based segmentation into graspable object cluster. Candidate grasps are then selected for these clusters to align the fingers centrally along the longer edges of the bounding box corresponding to each detected object. This grasp configuration was then used as the target pose for a task-space controller. This approach requires the extrinsic calibration of the camera with respect to the base of the arm.

3.2 Experimental Protocol

1. The method is evaluated using two experimental protocols.

2. In the first protocol, "with replacement", the gripper places any grasped objects back into the bin.

3. This tests the ability to grasp object in clutter, but also allows the robots to repeatedly pick up easy objects.

4. The second protocol is "without replacement".

5. The robots remove objects from the bin.

3.3 Important observations

1. Grasp success rate continues to improve as more data was accumulated.

2. A high success grasp, exceeding the baseline, was not achieved until at least halfway through the data collection process.

3. In the transfer experiment, they found that even a network trained on data from a different robot already performs better than chance, indicating a non-trivial amount of transfer.

4. A small number of objects account for a large number of failed grasps.

4 Questions

1. What is nonprehensile manipulation?