Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-16 (2016) 177–183

Autonomous Leaf Picking Using Deep Learning and Visual-Servoing

Konrad Ahlin * **Benjamin Joffe** ** **Ai-Ping Hu** ***
Gary McMurray **** **Nader Sadegh** †

* *Georgia Institute of Technology (e-mail: kahlin@gatech.edu)*
** *Georgia Institute of Technology (e-mail: joffe@gatech.edu)*
*** *Georgia Institute of Technology (e-mail: Ai-Ping.Hu@gtri.gatech.edu)*
**** *Georgia Institute of Technology (e-mail: Gary.McMurray@gtri.gatech.edu)*
† *Georgia Institute of Technology (e-mail: sadegh@gatech.edu)*

Abstract: Grasping unstructured objects with a robotic manipulator is a difficult task. The challenge lies in both determining which objects are desirable to grasp and where those objects are relative to the manipulator. However, a robotic arm with an eye-in-hand camera can be used to effectively grab and hold desired objects, even if the exact geometry and position of the objects are unknown before the experiment. Combining Convolutional Neural Networks in image processing and Monoscopic Depth Analysis for visual-servoing, it was successfully demonstrated that a robotic manipulator could be used to accurately locate and grasp a leaf from a plant while only using a single camera for identification. This task of finding and manipulating a leaf shows the potential of using robotic manipulators in increasingly unstructured environments.

Keywords: Deep Learning, Visual Servoing

1. INTRODUCTION

Robotic arms work best in structured environments where their desired tasks are known before they need to be executed. However, the field of robotics is ever expanding out of factory-like environments and into broader applications. Thus, the demand for completing unstructured tasks, tasks that have elements that cannot be known ahead of time, is increasing. This research will focus on the unstructured task of having a robotic arm grasping a leaf from a plant. The challenges in this task involve both the identification of a leaf within the space of a camera image as well as locating the leaf in the 3D Cartesian world. While the type of leaf can be assumed to be known, and the approximate location of the plant is assumed to be known, neither the exact geometry of the leaf nor the approximate location of the leaf is known. The task for the arm is to search for the leaf, identify the leaf, and then, using 2D pixel information, grab the leaf with its manipulator.

Identifying and localizing a leaf is not a trivial problem. Leaves vary in size and appearance, and are susceptible to overlapping and occlusions. Moreover, the implementation should be robust to variations in natural illumination. Deep Learning techniques have been shown to achieve great results in object detection, while demonstrating good computational performance by using modern GPU computing. Once information about the leaf can be determined in the 2D image space, the manipulator can attempt a routine to grasp the object.

Visual-servoing is the process of controlling a robotic device using real-time visual information in a feedback loop. Image Based Visual-Servoing (IBVS) is a classical approach to visual-servoing in robotics that attempts to converge on an object in 3D space by only using information about the objects 2D pixel geometry. IBVS does not assume any information about the object being viewed; instead, this method uses a given set of desired points in the image space that it uses for its error calculations [Chaumette and Hutchinson (2006)]. Image Based Visual-servoing is a widely studied topic that is often used in robotics applications.

Robotic arms typically have a virtue in their design that offer an alternative or supplement to the IBVS approach. Assuming that the kinematics of the arm and the rotation of the joint angles are known (a necessity for most joint control methods), the final position of the end-effector can be determined. Since the camera is rigidly attached to the manipulator, the position and orientation of the camera is always known in Cartesian space. Using the location of the camera and the pixel coordinates of desirable feature points, the 3D location of images can be determined. Monoscopic Depth Analysis (MDA) uses multiple images and triangulation to deduce the position of feature points in Cartesian space relative to the camera. Although the method of visual triangulation is widely known, the implications of using the MDA approach for robotic manipulators is potentially vast. It is a simple routine that has been used in other visual applications (eg. Kalghatgi and Sadegh (2012)) and holds promise in improving the control of robotic manipulators.

The task of picking a leaf can be separated into the following, unstructured sub-tasks: determining the position

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.10.033

of the leaf in the image space and transforming its position into Cartesian space. The method of using Convolutional Neural Networks as a means of identifying the leaf in the image space will be examined. To demonstrate the effectiveness of MDA as a control scheme, the typical elements of classical IBVS approach as well as MDA will be discussed for the purpose of comparison. This approach to visual-servoing was tested using an experimental setup and the results will be discussed.

2. IMAGE PROCESSING

2.1 Leaves detection

Object detection – a fundamental problem in Computer Vision – consists of producing the bounding boxes around the objects of desired class on an image. Traditional Machine Vision techniques are generally not suitable for the detection of objects like leaves. First, the leaves on the plants have complex backgrounds, that often include other leaves, which makes it hard to separate an individual leaf from the background. Second, leaves have complex shapes and appearances, thus creating a model and estimating the parameters is not computationally feasible for a real-time application. Finally, since the plants are located in the field there is a wide variation in natural illumination. That excludes many color-based approaches that have to be finely tuned to the lighting conditions. Hence, the problem of leaf detection requires using Machine Learning algorithms. In recent years Convolutional Neural Networks have become a popular approach.

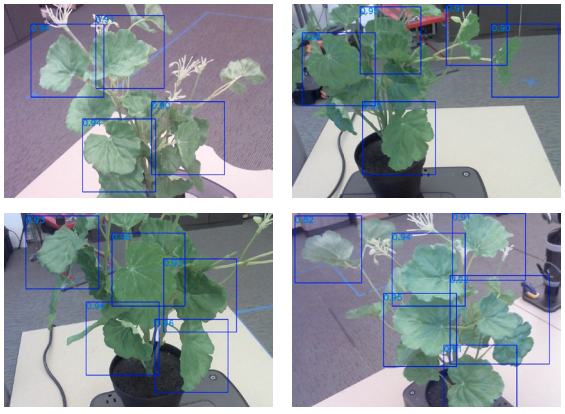

Fig. 1: Examples of leaves detections. The number in the corner of a bounding box indicates the score of the detected leaf.

To train the model we accumulated a small dataset of leaves images cropped to 227x227, labeled with two classes (leaves and background). Our network is based on the AlexNet architecture [Krizhevsky et al. (2012)] and consists of five convolutional layers and two fully-connected layers. The model was fine-tuned by initializing with weights trained on a large ILSVRC12 dataset (1.2 million images), and than trained on our task specific dataset of leaves. This common technique greatly improves the robustness and accuracy of the model. During deployment the algorithm treats fully-connected layers as convolutional and, thus, gets the probability map for the target classes at once, instead of making many classification calls in a traditional sliding window approach.

In our particular problem, we do not have to detect all the leaves in the frame, but rather identify a good candidate for subsequent sampling, i.e. the one in front of the camera, not occluded, and not having extreme angles that are hard for the robotic arm to approach. That filtering is incorporated in the dataset creation (where negatives include the leaves at extreme angles) and in candidate leaf processing (where after estimating 3D coordinates it can be dropped if deemed unsuitable).

2.2 Leaves tracking

To allow for successful execution of the leaf picking, there is a need to continuously track the target leaf from frame to frame as we move the robotic arm. Furthermore, triangulation of 3D coordinates requires precisely matched points within the target leaf for two consecutive frames. Since, during the execution, we have only one target leaf from the old frame and few candidate leaves in the new frame, we can efficiently obtain the SURF descriptors for these leaves (which would otherwise be computationally prohibitive). Then, to rule out the keypoints in the background we use green color thresholding on the leaf image; we also exclude the points in the corners, because they could belong to other leaves appearing within the bounding box.

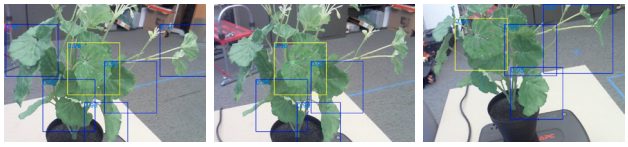

Fig. 2: Leaf tracking on several frames. Yellow bounding box indicates the target leaf that is being tracked; blue boxes indicate other detections in the frame; keypoints are highlighted on the leaves.

For each pair of the old target leaf and a new candidate leaf, we perform quick matching with FLANN. We filter out the outliers with RANSAC algorithm that estimates the largest set of points that agrees with a perspective transformation. At the end of this procedure the correct leaf has significantly higher number of matches and is updated as our target. If all the leaves have low number of matches, we assume that the target leaf is missing in the current frame. We continue searching for the leaf for an arbitrary number of frames, and if it is not found, we pick a new target.

2.3 Feature points

The approach discussed in the previous section typically produces over one hundred matched feature points within a leaf. The points, however, are consistent only between one pair of consecutive frames, which makes it not suitable for Image-Based Visual Servoing. Identifying consistent features on a leaf is difficult because for each frame a given leaf will change its position and orientation within the bounding box, and on some occasions parts of the leaf can be outside of the bounding box. The usage of bounding box regression [Weicheng Kuo (2015)] and semantic segmentation-aware models [Gidaris and Komodakis