International Conference of Agricultural Engineering
AgEng 2014 Zurich 6–10 July

Ref: C0137

Depth Based Fruit Detection from Viewer-Based Pose

Ehud Barnea and Ohad Ben-Shahar, Ben-Gurion University of the Negev

Abstract

Seeking to accurately detect and localize fruit of any color in 3D space for selective agrobotical operations, we exploit data given by Time-of-Flight or RGB-D cameras and propose a novel shape-based fruit detector using a fruit pose reference frame relative to the viewer. Surface normals, which are shaped-based local features, are accumulated into bins of different shapes along the reference frame's axes. The used normals are represented using two angles from a viewer-based reference frame, to achieve a representation that is suitable for fruit types that are almost symmetric around an axis (e.g., bell-peppers), without having an effect on fruit types with no axis-symmetry. Results are shown on a particularly challenging pepper dataset.

Keywords: Fruit Detection, Harvesting Robots

1 Background

The recent advances in the production of depth cameras have made it easy to acquire depth information of scenes (both indoors and outdoors) in the form of RGB-D images or 3D point clouds. The depth modality, which is inherently different than color and intensity, is lately being employed to solve many kinds of general computer vision problems, such as object recognition (Lai et al., 2011), object detection (Hinterstoisser et al.,2011; Spinello and Arras, 2011), pose estimation (Shotton et al.,2013;Aldoma et al.,2011) and segmentation (Silberman and Fergus,2011). Seeking to improve agricultural robotics, RGB-D images were also used to detect several kinds of fruit (Chi and Ling, 2004; Edan et al., 2000; Hannan and Burks, 2004). The depth modality may be especially suited for agricultural application, as the shape of visible objects becomes explicit and unaffected by lighting conditions, which tend to vary greatly in several agricultural applications (Harrell et al., 1989). Furthermore, as fruit sometimes have color similar to the surrounding, shape becomes a more prominent feature allowing us to separate fruit from background.

Inspired by recent advancements in general computer vision and previously suggested fruit detection systems (Kapach et al., 2012), our proposed detector attempts to exploit the shape of the fruit using a reference frame directed to the viewing position and construct a representation that is based on this reference frame. More specifically, we construct a feature vector by processing surface normal angles accumulated into histograms of the bins they fall in, which are calculated relative to that frame. In what follows we discuss the relevant background, followed by an elaborate description of our suggested algorithm and display detection results over a challenging bell-pepper dataset.

1.1 Employing the depth modality

The depth modality has been applied to most kinds of computer vision problems, even more so following the release of the Kinect camera. This kind of data, in contrast to plain RGB, enables a much easier calculation of various things such as the separation of foreground from background (by thresholding distances), floor detection and removal (by finding the dominant plane in the scene (Aldoma et al.,2011)), segmentation of non-touching objects (by identifying and removing the floor (Aldoma et al.,2011)), or the estimation of surface normals and curvature (e.g., by fitting planes to 3D points in small neighborhoods (Rusu, 2010)).

In previous studies, depth images were often treated as intensity images (Spinello and Arras, 2011; Janoch et al., 2011), allowing them to be used with previous algorithms requiring regular images. Exploiting surface *normals* as well, Hinterstoisser et al. (Hinterstoisser et al., 2012) recently suggested a combined similarity measure by considering both image gradients and surface normals. This was done to combine the qualities of color with depth, as strong edges are prominent mostly *on* object silhouettes, whereas the normals make explicit the shape *between* silhouettes.

1.2 RGB-D object and fruit detection

Object detection in color images has been a subject of research for many years. With the introduction of depth data to the field new challenges emerge, such as how to properly use this kind of data, or how it may be used in conjunction with RGB data to combine the spatial characteristics of both channels. See Figure 1 for an example of a registered pair of color image and depth image taken from our pepper dataset.

In a recent indoor RGB-D object detection dataset presented by Janoch et al. (Janoch et al., 2011), a baseline was suggested employing the popular part-based detector by Felzenszwalb et al. (Felzenszwalb and McAllester, 2008). Taking object's parts into account, this model is a variant of the HoG algorithm (Dalal and Triggs, 2005) that combines a sliding window approach together with a representation based on histograms of edge orientations. This detector was employed by running it on depth images while treating them as intensity images. The results indicated that applying this algorithm to color images always gives better results than applying it over depth images, which can be seen to suggest that the depth information should be regarded differently than color. Tang et al. (Tang et al., 2012) also used the HoG formulation, but with histograms of surface normals that are characterized by two spherical angles.

Apart from object detection research, more than a few *recognition* algorithms have been suggested. Due to the different nature of the problem, only one object is visible and sometimes segmentation is assumed. Seeking to employ such algorithms for the problem of detection, Kim et al. (Kim et al., 2013) suggested examining a small set of segmentation hypotheses. A part-based model generalized to 3D is used together with HoG features and other 3D features of the segmented objects. This scheme results in a feature vector containing both color features and 3D features, which gives better results in most categories.

Depth information was also previously utilized to detect specific types of fruit. An algorithm by Jimenez et al. (Jimenez et al., 2000) exploits the spherical shape of oranges by looking for and aggregating a set of local primitive that are likely to belong to spherical objects. Monta and Namba (Monta and Namba, 2003) on the other hand, combined between the depth and color by improving a standard color algorithm with the addition of depth in order detect tomatoes. They start by detecting red pixels, which easily separates tomato bunches from the background, and then depth is employed to distinguish between specific tomatoes.

2 Feature vector construction

We incorporate our design of feature vectors into the well-known “sliding window” principle. A given depth image is represented as a 3D point cloud (a set of points in 3D), and is efficiently scanned with a fixed-sized box for which the existence of fruit is examined. To do so, a feature vector is constructed using the points inside a box, and a SVM machine-learning classifier is used to classify the vector as either representing fruit or not. Since a single fruit is usually present in several boxes, very close boxes that are found to contain fruit are merges using the mean-shift algorithm. In the following subsections we will describe the feature vector construction process.

2.1 Reference frame calculation

By going over every possible box in space, we are guaranteed that every fruit will coincide with a box such that the center of the fruit and the center of the box will be very close. Therefore, a reference frame based in the center of the box can be regarded as a frame in the center of a fruit. More formally, for a given box with center point , seen from viewpoint , we construct a frame based at point from vectors defined as:

Examining different fruit translations, when the position of a fruit varies so as the visible portion. The defined reference frame is constructed such that vector will always go through the visible portion of the fruit. While the choice of viewer-based or fixed is irrelevant for non-axial-symmetric fruit, it allows us to construct a partially translation invariant feature vector when applied to axial-symmetric fruit, as will be clarified below.

2.2 Binning scheme

Points inside the 3D box are distributed into a 2D array of equally sized bins according to distances between each point and the reference axes. More specifically, a point is placed in a bin according to two distances - Euclidean distance from the point and the plane spanned by vectors and (passing through), and angular distance between the vector (connecting the box's center to the point) and the vector .
Following the placement of points to bins, the surface normal at each point is calculated for a small neighborhood around the point by fitting a plane to neighboring points (the normal of which is an estimation of the surface normal at that point). The estimated normals of all the points in each bin are accumulated into two 1D angle histograms. The first histogram is for the angles between surface normal and the vector and the second is for angles between surface normal and the vector (connecting the box's center to the point). The histograms of all the bins are concatenated to form a vector, which we can then use for SVM training or classification.
Since one of the angles in the representation of normals is defined relative to the vector, and depends only on horizontal translations, the angle from is not affected by horizontal translations when an axis-symmetric fruit is roughly positioned such that the axis is pointing upwards.

3 Experimental results

We evaluate the results of our suggested approach on a dataset of 100 images taken at a greenhouse and containing both red and green bell-peppers together with 3D ground truth annotations of fruit boxes. Our detector is executed over all the dataset images, and after a short pruning phase that disregards dark areas with no highlights, it goes on to seek both red

and green peppers. Since the dataset contains highly occluded peppers, it is tested only on those that are at least 50% visible, and a suggested detection is considered true when the suggested box overlaps a ground truth box (in the case of several suggested boxes overlapping the ground truth only one suggestion is deemed correct). The results are depicted in Figure 2, in which the precision and recall curve is displayed, so for a specific system one may choose the wanted relation between precision and recall.

4 Conclusions

We have presented a fruit detector based on local 3D shape features calculated relative to a viewer-based reference frame. The proposed detector is able to detect the 3D location of fruit regardless of color and is partially invariant to translations under certain conditions. Good results were achieved over a challenging dataset of both green and red bell-peppers, showing that automatically picking such fruit is possible. A main direction for future research is the incorporation of color. While the usage of color may not help much when all peppers are green, when dealing with a combination of both green and red fruit it may lead to better results should the color of non-green fruit is incorporated into the detection process.

5 Acknowledgements

This research was funded in part by the European Commission in the 7th Framework Programme (CROPS GA no. 246252), partially supported by the Helmsley Charitable Trust through the Agricultural, Biological and Cognitive Robotics Center of Ben-Gurion University of the Negev and the Israeli Ministry of Agriculture. The authors also thank the generous support of the Frankel fund.

6 References

Aldoma, A., Vincze, M., Blodow, N., Gossow, D., Gedikli, S., Rusu, R. B., & Bradski, G. (2011). CAD-model recognition and 6DOF pose estimation using 3D cues. In *Computer Vision Workshops (ICCV Workshops)*, 2011 IEEE International Conference on (pp. 585-592).

Chi, Y. T., & Ling, P. P. (2004). Fast fruit identification for robotic tomato picker. In *ASABE Annual International Meeting*.

Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005*. (Vol. 1, pp. 886-893).

Edan, Y., Rogozin, D., Flash, T., & Miles, G. E. (2000). Robotic melon harvesting. *Robotics and Automation, IEEE Transactions on*, 16(6), 831-835.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008*. IEEE Conference on (pp. 1-8).

Hannan, M. W., & Burks, T. F. (2004). Current developments in automated citrus harvesting. *ASAE* paper, 43087.

Harrell, R. C., Slaughter, D. C., & Adsit, P. D. (1989). A fruit-tracking system for robotic harvesting. *Machine Vision and Application*

Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., & Lepetit, V. (2011). Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *Computer Vision (ICCV), 2011* IEEE International Conference on (pp. 858-865).

Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., & Lepetit, V. (2012). Gradient response maps for real-time detection of textureless objects. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, 34(5), 876-888.

Janoch, A., Karayev, S., Jia, Y., Barron, J. T., Fritz, M., Saenko, K., & Darrell, T. (2013). A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision* (pp. 141-165).

Jiménez, A. R., Ceres, R., & Pons, J. L. (2000). A vision system based on a laser range-finder applied to robotic fruit harvesting. *Machine Vision and Applications*, 11(6), 321-329.

Kim, B. S., Xu, S., & Savarese, S. (2013). Accurate Localization of 3D Objects from RGB-D Data Using Segmentation Hypotheses. In *Computer Vision and Pattern Recognition (CVPR), 2013* IEEE Conference on (pp. 3182-3189).
Lai, K., Bo, L., Ren, X., & Fox, D. (2011). Sparse distance learning for object recognition combining rgb and depth information. In *Robotics and Automation (ICRA), 2011* IEEE International Conference on (pp. 4007-4013).
Monta, M., & Namba, K. (2003). Three-dimensional sensing system for agricultural robots. In *Advanced Intelligent Mechatronics, 2003*. Proceedings. 2003 IEEE/ASME International Conference on (Vol. 2, pp. 1216-1221).
Rusu, R. B. (2010). Semantic 3D object maps for everyday manipulation in human living environments. Ph.D Thesis, KI-Künstliche Intelligenz, 24(4), 345-348.
Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1), 116-124.
Silberman, N., & Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011* IEEE International Conference on (pp. 601-608).
Spinello, L., & Arras, K. O. (2011). People detection in rgb-d data. In *Intelligent Robots and Systems (IROS), 2011* IEEE/RSJ International Conference on (pp. 3838-3843).
Tang, S., Wang, X., Lv, X., Han, T. X., Keller, J., He, Z.,& Lao, S. (2013). Histogram of oriented normal vectors for object recognition with a depth sensor. In *Asian Conference on Computer Vision 2012* (pp. 525-538).

Figure 1: A pair of registered color and depth image pair. Darker depth represents a shorter distance from the camera

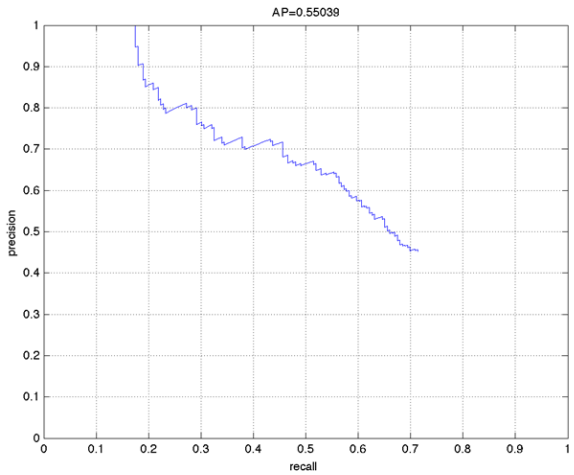

Figure 2: Precision and recall curve over the entire pepper dataset