TOWARDS AUTOMATED CROP YIELD ESTIMATION:
Detection and 3D Reconstruction of Pineapples in Video Sequences

Supawadee Chaivivatrakul, Jednipat Moonrinta, Matthew N. Dailey
School of Engineering and Technology, Asian Institute of Technology, Pathumthani Thailand
supawadee.chaivivatrakul@ait.ac.th, jednipat.moonrinta@ait.ac.th, mdailey@ait.ac.th

Keywords: Object detection, Keypoint detection, Keypoint descriptors, Keypoint classification, Image segmentation, Structure from motion, 3D reconstruction, Ellipsoid estimation, Pineapple, Mobile field robot, Agricultural automation

Abstract: Towards automation of crop yield estimation for pineapple fields, we present a method for detection and 3D reconstruction of pineapples from a video sequence acquired, for example, by a mobile field robot. The detection process incorporates the Harris corner detector, the SIFT keypoint descriptor, and keypoint classification using a SVM. The 3D reconstruction process incorporates structure from motion to obtain a 3D point cloud representing patches of the fruit's surface followed by least squares estimation of the quadric (in this case an ellipsoid) best fitting the 3D point cloud. We performed three experiments to establish the feasibility of the method. Experiments 1 and 2 tested the performance of the Harris, SIFT, and SVM method on indoor and outdoor data. The method achieved a keypoint classification accuracy of 87.79% on indoor data and 76.81% on outdoor data, against base rates of 81.42% and 53.83%, respectively. In Experiment 3, we performed 3D reconstruction from indoor data. The method achieved an average of 34.96% error estimating the ratio of the fruits' major axis to short axis length. Future work will focus on increasing the robustness and accuracy of the 3D reconstruction method as well as resolving the 3D scale ambiguity.

1 INTRODUCTION

Agricultural automation has the potential to improve farm yields, improve crop quality, and lower production costs. In particular, autonomous in-field inspection of fruit fields could improve farmers' ability to track crops over time, plan maintenance and harvesting activities, and predict yield. We are interested in developing autonomous inspection robots for pineapple farms that use low-cost cameras and machine vision to isolate and grade the fruit while it is still in the field. We focus here on the related problems of 1) segmenting a video to find the pineapple fruit in the field, and 2) obtaining 3D models of detected fruits.

In segmentation, the first main challenge is that since the plants are tightly spaced, we cannot typically see all of a particular fruit, ruling out shape-based methods. Second, since pineapples often have a similar color to the rest of the plant, we cannot rely on color. We thus rely on texture. We use the fast Harris algorithm (Harris and Stephens, 1988) to find corner points then apply the SIFT descriptor (Lowe, 2004) to those detected points. We then classify the descriptors using a SVM (Support Vector Machine).

For 3D modeling, we combine the pineapple detector with 3D point cloud estimation using well known structure from motion techniques (see, e.g., Pollefeys et al., 2004), and then we perform least-squares estimation of the quadric (in this case an ellipsoid) best fitting the 3D point cloud. Our current method is not robust to outlier points and does not resolve the scale ambiguity of the 3D reconstruction (future work will address these limitation), but the method does provide useful information about fruit orientation and shape.

The experiments establish the feasibility of using texture to segment pineapples in video sequences and using structure from motion to reconstruct pineapple shape. This work is a step towards fully automatic crop yield by mobile field robots.

2 METHODOLOGY

Our methodology consists of frame selection, image segmentation, and 3D reconstruction from the point cloud. In our current prototype, we select two views of each fruit in a video sequence manually. However, we plan to perform automated key frame selection in future work. We detail each of the other steps in turn.

2.1 Image Segmentation

Our segmentation algorithm consists of keypoint extraction, keypoint descriptor calculation, keypoint classification, and morphological operations to retrieve the fruit region in a given image.

We use the Harris corner detector (Harris and Stephens, 1988) to find candidate keypoints over the whole image, since images of a pineapple's surface have many points with corner-like structure. We find that in practice, the Harris detector tends to find fairly dense sets of keypoints on pineapple image regions that are very useful for reconstructing 3D point clouds representing the fruit surface.

Classifying the keypoints as pineapple and non-pineapple points requires a rich description of the local texture surrounding the keypoint. We compute SIFT descriptors (Lowe, 2004) (a 128-element vector) for the high-gradient Harris keypoints that are not too close to image boundaries.

We use support vector machines (SVM) to classify keypoints as pineapple or non-pineapple. In other work, we have performed experiments on SIFT keypoint descriptor classfication using a variety of SVM kernels and hyperparameter settings, and we find that the radial basis function (RBF) kernel has the best overall performance. Here we use RBFs with a cross-validated grid search over hyperparameter settings to find the best hyperparameter settings.

For segmentation, we find contiguous pineapple regions using morphological closing to connect nearby pineapple points then remove regions smaller than 25% of the expected fruit area, based on assumptions of image resolution and distance to the camera.

2.2 3D Reconstruction

To obtain 3D point clouds from candidate pineapple image regions, we find point correspondences between image pairs and then apply standard algorithms from the structure from motion literature, as described in the following sections.

The first step is feature point extraction. Once pineapple regions have been identified in a pair of images of the same fruit, we extract SURF (Bay et al., 2008) feature points from those regions. Although the Harris corner detector and the SIFT keypoint descriptor work well for image segmentation, we find that the standard SURF algorithm gives us more reliable correspondences for 3D point cloud reconstruction.

To find point correspondences between two images, we find, for each keypoint descriptor in one image, the most similar descriptor in the other image. We use the dot product similarity measure with a threshold to find the most likely corresponding keypoint in one image for each keypoint in the other.

To remove outliers in the resulting set of putative correspondences, we use the adaptive RANSAC method for fundamental matrix estimation (Hartley and Zisserman, 2004) to find the best fundamental matrix and correspondence consensus set, removing outliers inconsistent with the epipolar geometry. The remaining inlier points are used for 3D point cloud estimation.

The next step is 3D point cloud estimation. We assume that the camera's intrinsic parameters are fixed and given as a calibration matrix $\mathtt{K}$. We next estimate camera matrices for the two images, using the essential matrix method (Hartley and Zisserman, 2004).

Once two camera matrices are known, we compute linear estimates of all of the 3D points then refine those estimates using nonlinear least squares (Levenberg-Marquardt).

In a real field, we cannot rotate the pineapple or move the camera to get a complete view of the fruit. Therefore, we must estimate the fruit's shape from a partial view. We propose an algorithm for reconstructing the 3D shape of a pineapple from a 3D point cloud estimated from a partial view of the fruit's surface.

Since pineapples are nearly ellipsoidal, we model each fruit as an ellipsoid and perform least squares estimation of the ellipsoid's parameters to fit the point cloud data estimated in the previous step. Using Li and Griffiths' (2004) method, we actually estimate the quadric

$$\mathtt{Q} = \begin{bmatrix} a & h & g & p \\ h & b & f & q \\ g & f & c & r \\ p & q & r & d \end{bmatrix} \tag{1}$$

defining $\mathbf{X}^T \mathtt{Q} \mathbf{X} = 0$ using least squares.

Once the best-fitting ellipsoidal quadric $\mathtt{Q}$ is found, we extract the ellipsoid's center, orientation, and axis radii.

3 EXPERIMENTAL RESULTS

To evaluate our methods, we performed three experiments: fruit segmentation on indoor data, fruit

Table 1: Distribution of training and test keypoints for indoor and outdoor segmentation.

Data set	Number of Fruit	Positive Instances	
		Indoor	Outdoor
Training	20	88.76%	35.48%
Test	10	81.42%	46.17%

segmentation on outdoor data, and 3D fruit reconstruction. The 3D reconstruction experiment was only applied to indoor data.

3.1 Segmentation

In Experiment 1, we captured indoor videos of 30 pineapples from a distance of approximately 30 centimeters. We chose one frame from each video and split the data into 20 training and 10 test images. For every image, we applied the Kovesi implementation of the Harris corner detector (Kovesi, 2000) with a Gaussian smoothing standard deviation of 0.5, a threshold of 1, and a non-maximum suppression radius of 2. We then extracted SIFT descriptors for the Harris corner points by modifying an open implementation of SIFT (Vedaldi, 2006) using 4 scales and 8 orientations. The distribution of positive (pineapple) and negative (background) keypoints over the training set and test set is shown in Table 1.

We then built SVM models with the RBF kernel using LIBSVM (Chang and Lin, 2001). The RBF kernel based SVM requires two hyperparameters, c, which controls the tradeoff between training error and model complexity, and γ, which controls the width of the RBF kernel. We used a grid search and 5-fold cross validation within the training set to find optimal values of c and γ then used the best parameters to train a final model on all of the training data then used the resulting model to classify the test set.

The best parameter setting for the indoor data was $(c = 2^2, \gamma = 2^0)$. Accuracy data for the classifiers are shown in Table 2, and a sample of the results is shown in Figure 1.

In Experiment 2, we performed the same steps on data acquired outdoors. In the outdoor data, there were many more negative keypoints due to the complex background and occlusions (Table 1). The best parameter setting for the outdoor data was $(c = 2^{1.5}, \gamma = 2^{1.75})$.

After classifying points as lying on the pineapple surface, we performed morphological closing with a disk-shaped structuring element of radius 30 using Matlab's image processing toolbox (The Mathworks, 2007). A sample of the results is shown in Figure 2(a) We performed further processing on connected regions containing 20,000 or more pixels (at 30 cm, with our camera, the visible pineapple surface region typically contains approximately 80,000 pixels).

Table 2: SVM keypoint classification accuracy for indoor and outdoor segmentation.

Model type	Data set	Accuracy	
		Indoor	Outdoor
Cross-validation	Training	97.76%	99.85%
	Validation	93.86%	82.64%
Final	Training	97.67%	99.80%
	Test	87.79%	76.81%

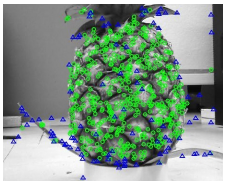
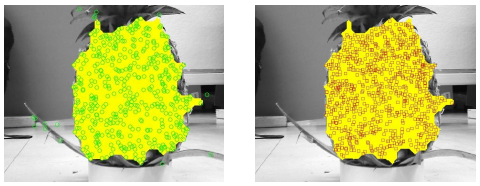


Figure 1: Sample SVM classification results, the green circle is pineapple point.

3.2 3D Reconstruction

In Experiment 3, we next applied Strandmark's (2008) Matlab port of the SURF reference implementation to the pineapple region detected in each image. A sample of the feature points we obtained is shown in Figure 2(b). Sample results from our structure from motion stream (matching and outlier detection, point cloud estimation, a Delaunay triangulation of the point cloud, and an estimated ellipsoid) are shown in Figures 3–4. A quantitative evaluation of the ellipsoid estimation is presented in Table 3. We found that the outdoor pineapple data had too many outlier points for accurate ellipsoid estimation, so the data in Table 3 are only for our indoor data.



(a) (b)

Figure 2: Pineapple segmentation and SURF feature points. (a) Result of closing operation to obtain pineapple region. (b) SURF feature points in the pineapple region.

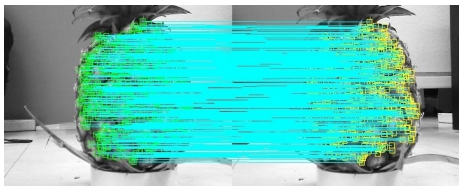
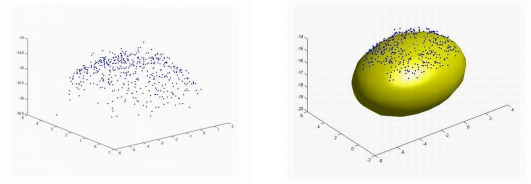


Figure 3: Feature point matching after outlier removal with RANSAC.



(a) (b)

Figure 4: Sample 3D reconstruction results. (a) 3D point cloud reconstructed from the inlier points in Figure 3. (b) Estimated ellipsoid.

4 CONCLUSION AND DISCUSSION

Our experiments demonstrate the feasibility of the proposed methods. The SVM model for pineapple keypoint classification achieves an accuracy of 87.79% for indoor data, which is sufficient, with post processing, to accurately segment pineapple and non-pineapple regions. The SURF keypoints computed for these regions enable 3D point cloud estimation with sufficient accuracy in most cases to estimate an ellipsoid with roughly accurate dimensions. In future work, we will focus on increasing the robustness of the method and resolving the 3D scale ambiguity, and perform experiments with real field robots.

Table 3: Ratios of medium axis length to major axis length and short axis length to major axis length for actual fruit and estimated ellipsoids.

No.	Fruit		Ellipsoid		Error	
	medium/ major	short/ major	medium/ major	short/ major	medium/ major	short/ major
1	0.83	0.83	0.71	0.58	14.46%	30.62%
2	0.96	0.96	0.66	0.58	31.54%	40.07%
3	0.85	0.85	0.65	0.53	23.41%	37.28%
4	0.84	0.84	0.69	0.57	18.66%	32.69%
5	0.88	0.88	0.73	0.58	17.04%	34.46%
6	0.85	0.85	0.72	0.58	15.57%	31.92%
7	0.86	0.86	0.63	0.53	26.30%	38.17%
8	0.89	0.89	0.71	0.60	19.99%	32.20%
9	0.83	0.83	0.67	0.56	19.39%	32.88%
10	0.88	0.88	0.65	0.54	25.96%	39.34%
Average	0.87	0.87	0.68	0.56	21.23%	34.96%

ACKNOWLEDGEMENTS

SC was supported by a graduate fellowship from the Thailand National Science and Technology Development Agency (NSTDA). Apisit Aroonnual helped develop the ellipsoid fitting software. We thank Anupun Terdwongworakul, Paul Janecek, and the members of the AIT Vision and Graphics Lab for useful comments on the work.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110:346–359.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition.

Kovesi, P. (2000). MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering,The University of Western Australia. Available at `http://www.csse.uwa.edu.au/~pk/research/matlabfns/`.

Li, Q. and Griffiths, J. G. (2004). Least squares ellipsoid specific fitting. In *Proceedings of the Geometric Modeling and Processing Conference*. IEEE Computer Society.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110.

Pollefeys, M., van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R. (2004). Visual modeling with a hand-held camera. *International Journal of Computer Vision*, V59.

Strandmark, P. (2008). SURFmex [open source software]. Available at `http://www.maths.lth.se/matematiklth/personal/petter/surfmex.php`.

The Mathworks (2007). Image processing toolbox user's guide. Technical report.

Vedaldi, A. (2006). An implementation of SIFT detector and descriptor. Available at `http://www.vlfeat.org/~vedaldi/assets/sift/sift.pdf`.