Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-16 (2016) 090–095

Multi-camera Fruit Localization in Robotic Harvesting

S. S. Mehta * **T. F. Burks** **

* *University of Florida, Department of Industrial and Systems Engineering, Shalimar, FL-32579 (e-mail: siddhart@ufl.edu).*
** *University of Florida, Department of Agricultural and Biological Engineering, Gainesville, FL-32611 (e-mail: tburks@ufl.edu)*

Abstract: Motivated by an effort to study *layered vision systems* for robotic harvesting, this paper investigates the problem of fruit localization using multiple cameras in the fruit detection layer of the vision system. A pseudo stereo-vision approach is presented where fruit matching is accomplished by loosely holding the epipolar constraint to reduce computation time. In the presence of noise, heuristics are presented to identify the greatest subset of cameras with matched fruits. Subsequently, the fruit depth is obtained by minimizing the summation of the image reprojection error in cameras with matching fruit. Monte Carlo simulations are performed to establish localization efficiency of the proposed approach under varying design parameters and image noise.

Keywords: Fruit localization, fruit matching, robotic harvesting, pseudo stereo-vision

1. INTRODUCTION

Vision systems are widely accepted in robotic harvesting due to their ability to provide information rich image feedback of the environment. An image is formed by projecting the three-dimensional (3D) Euclidean space onto a two-dimensional image plane. This projection results in loss of depth information. It is well known that various operations such as manipulator path planning and servo control can significantly benefit from knowing the fruit depth or alternately the 3D fruit position. For example, given a global map of 3D fruit positions, a path planning algorithm can find the shortest path for a manipulator to harvest fruits thereby reducing harvesting time. The process to obtain 3D fruit position by recovering or measuring the depth information is referred to as fruit localization.

Approaches to localization can be broadly classified into systems using vision (or cameras) as the only sensor and systems using range measuring devices in conjunction with vision. The later class of systems may include a laser range finder or an ultrasound transducer to determine the fruit depth by measuring the time-of-flight, i.e., the time required for the light or sound to travel to the fruit and back (e.g., see Harrell et al. [1990], Ceres et al. [1998], Jiménez et al. [2000], Bulanon and Kataoka [2010]). Time-of-flight based cameras use similar principle as laser range finders to generate 3D maps. Due to high speeds and less complexity, they are promising in agricultural applications (see Karkee et al. [2014]). However, the use of time-of-flight cameras is limited due to high cost and low resolutions. The vision-only fruit localization systems are based on machine vision principles to determine the unknown depth. Structure-from-motion (Muscato et al. [2005]) and model-based (Mehta and Burks [2014]) approaches can be used for depth identification with single monocular camera systems. Stereo-vision or triangulation based approaches use the notion of disparity (i.e, the difference in image location of the same 3D point viewed by two cameras) to determine the depth. Despite added complexity and processing times, stereo-vision is widely used in fruit localization (e.g., see Buemi et al. [1996], Kondo et al. [1996], Recce et al. [1996], Van Henten et al. [2002, 2003], Font et al. [2014]). Moreover, with increase in processing power and the advent of parallel architectures that use FPGAs and GPUs, stereo-vision-like approaches hold significant potential. A comprehensive review of fruit localization methods can be found in Gongal et al. [2015].

We are investigating layered vision systems for robotic harvesting, where each layer may contain multiple monocular cameras. Specific functions can be assigned to each layer, for example, the layer farthest from the tree canopy can be designed for fruit detection and localization. Motivated by this sensing approach, we propose a multi-camera fruit localization approach where the number of cameras can be more than two, i.e., beyond standard stereo-vision. The fruit detection problem is assumed to have been solved (e.g., Plebe and Grasso [2001], Hannan et al. [2009]) to yield the fruit centroids. Given the centroids, fruit matching between cameras is accomplished using an epipolar constraint. The constraint is loosely applied which may introduce localization errors (false positives), however with multiple cameras (> 2) this adverse effect is reduced. Additionally, heuristics are defined to enable faster matching and yield the greatest set of "matched" cameras for each fruit in the presence of image noise. A pseudo stereo-vision

⋆ This research is supported by a grant from the USDA NIFA AFRI National Robotics Initiative #2013-67021-21074. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency.

10.1016/j.ifacol.2016.10.017

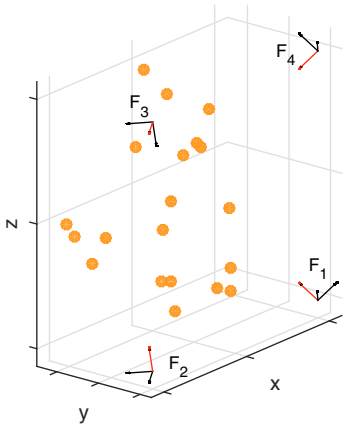

Fig. 1. The problem scenario showing multiple monocular cameras viewing randomly distributed fruits.

approach is taken to obtain fruit depths by minimizing the summation of the image reprojection error in the matched cameras. Monte Carlo simulations are performed to study the performance in terms of false positives, false negatives, localization error, and processing time when the design parameter and image noise are varied.

2. PROBLEM STATEMENT

Consider a scenario where multiple cameras possibly from different layers are viewing fruits as shown in Fig. 1. The system may include stationary (e.g., fixed to the workplace) and moving (e.g., attached to the robot's end-effector) cameras. Let $\mathcal{F}_i$ for $i = 1, 2, \ldots, n$ be the coordinate frames attached to $n \in \mathbb{Z}_{\geq 2}$ cameras that view $m \in \mathbb{Z}_{>0}$ stationary fruits. The cameras are calibrated such that the rotation and translation between the cameras is known. Various image processing methods (e.g., Plebe and Grasso [2001], Hannan et al. [2009]) can be used to obtain centroids of the visible fruits in the projected images. The homogeneous image coordinates of the fruit centroid be denoted by $p_{ij} \in \mathbb{R}^3$, where $j = 1, 2, \ldots, m$, and the corresponding unknown Euclidean position of the fruit in $\mathcal{F}_i$ be $\bar{m}_{ij} = [x_{ij}\ y_{ij}\ z_{ij}]^T$. Given the intrinsic camera calibration matrix A_i is known, the normalized Euclidean coordinates of $\bar{m}_{ij}$ can be obtained as $m_{ij} = A_i^{-1} p_{ij}$.

The objective is twofold. Firstly, the fruit correspondence between the cameras must be established to obtain matching fruit. Given the matching fruit, the second objective is to determine the fruit depth. For notational simplicity and without loss of generality, the subscript j from the above variables is eliminated to localization each fruit separately.

3. FRUIT MATCHING

Property 1. Epipolar constraint: As shown in Fig. 2, let the intersection of camera baseline with the image planes, i.e., epipoles, be e_1 and e_2. Given the image coordinates p_1 in the first camera, it can be seen that the fruit projection in the second camera must lie on the epipolar line l_2. Also, the image projection in the first camera must lie on l_1.

Consider $\mathcal{F}_1$ as the reference camera, and the objective is to determine fruit correspondence for a fruit with pixel coordinates p_1 and normalized Euclidean coordinates m_1.

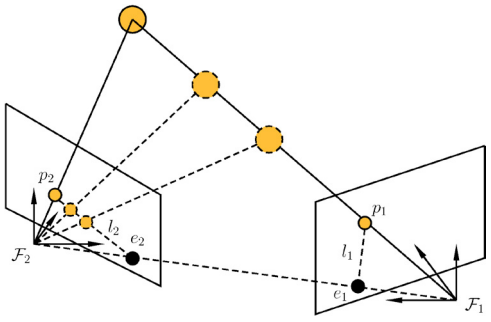

Fig. 2. Epipolar geometry for a pair of cameras viewing a fruit.

Given p_1, m_1, and the known relative position and orientation of the cameras, epipolar lines $l_2, l_3, \ldots, l_n$ in cameras $2, 3, \ldots, n$ can be obtained (see Hartley and Zisserman [2004] for details). From Property 1, the image coordinates of the target fruit in other cameras, i.e., $p_2, p_3, \ldots, p_n$, must lie on the corresponding epipolar lines. In practice, vision-based systems are subject to noise. The major source of noise during image acquisition is the sensor noise caused by poor illumination and high temperature, analog-to-digital conversion, and quantization. Image processing such as feature point tracking or centroid tracking, introduces additional pixel errors which collectively can be called as the image noise. Due to noise, the fruit may not appear exactly on the epipolar line. Another challenge is that from m fruits there can be more than one fruit close to the epipolar line in each camera. Among these candidate fruits, the goal is to determine a matching fruit to p_1 in cameras $2, 3, \ldots, n$. One solution is to obtain epipolar lines in the first camera for all candidate fruits from each camera, and p_1 must lie close to the epipolar line of the matching fruit for each camera. By minimizing the distance of the epipolar line from p_1, the fruit correspondence can be established. However, the computational burden on the system increases with the number of candidate fruits and cameras. Motivated to reduce processing time, a simple heuristic approach is presented to the fruit correspondence problem.

The candidate fruits with possible correspondence with p_1 in each camera are defined as those which lie within distance d from the epipolar line. The parameter d can be selected based on the prior knowledge of the image noise. If selected too small then no candidate fruit may be found. The effect of parameter d on the localization performance is analyzed in Section 5. Let s_i for $i = 1, 2, \ldots, n$ denote the set of candidate fruits for each camera. Note that when camera 1 is the reference camera $s_1 = p_1$. In addition, let $\#s_i$ denote the cardinality of s_i. Compared to image resolution, fruits appear relatively sparse in the image plane, hence $\#s_i \in \mathbb{Z}_{\geq 0}$. For the chosen d, the cardinality $\#s_i$ tells us whether unique fruit correspondence exists for camera i. If it exists, the camera i will be called a matched camera. Let $c_r \subset \{1, 2, \ldots, n\}$ be a set of matched cameras for the reference camera r. Given the candidate set s_i, the set c_r is obtained as

$$\#s_i = \begin{cases} 0 & \text{then } i \notin c_r \\ 1 & \text{then } i \in c_r \\ > 1 & \text{then } i \notin c_r \end{cases} \quad \text{for } i = 1, 2, \ldots, n. \qquad (1)$$