GRAPE YIELD ANALYSIS WITH 3D CAMERAS AND ULTRASONIC PHASED ARRAYS

A thesis by publications presented in fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Engineering

at Massey University, Albany,

New Zealand.

Baden Parr

2024

Abstract

Accurate and timely estimation of vineyard yield is crucial for the profitability of vineyards. It enables better management of vineyard logistics, precise application of inputs, and optimization of grape quality at harvest for higher returns. However, the traditional manual process of yield estimation is prone to errors and subjectivity. Additionally, the financial burden of this manual process often leads to inadequate sampling, potentially resulting in sub-optimal insights for vineyard management. As such, there is a growing interest in automating yield estimation using computer vision techniques and novel applications of technologies such as ultrasound.

Computer vision has seen significant use in viticulture. Current state-of-the-art 2D approaches, powered by advanced object detection models, can accurately identify grape bunches and individual grapes. However, these methods are limited by the physical constraints of the vineyard environment. Challenges such as occlusions caused by foliage, estimating the hidden parts of grape bunches, and determining berry sizes and distributions still lack clear solutions.

Capturing 3D information about the spatial size and position of grape berries has been presented as the next step towards addressing these issues. By using 3D information, the size of individual grapes can be estimated, the surface curvature of berries can be used as identifying features, and the position of grape bunches with respect to occlusions can be used to compute alternative perspectives or estimate occlusion ratios. Researchers have demonstrated some of this value with 3D information captured through traditional means, such as photogrammetry and lab-based laser scanners. However, these face challenges in real-world environments due to processing time and cost.

Efficiently capturing 3D information is a rapidly evolving field, with recent advancements in real-time 3D camera technologies being a significant driver. This thesis presents a comprehensive analysis of the performance of available 3D camera technologies for grape yield estimation. Of the technologies tested, we determined that individual berries and concave details between neighbouring grapes were better represented by time-of-flight based technologies. Furthermore, they worked well regardless of ambient lighting conditions, including direct sunlight. However,

distortions of individual grapes were observed in both ToF and LiDAR 3D scans. This is due to subsurface scattering of the emitted light entering the grapes before returning, changing the propagation time and by extension the measured distance. We exploit these distortions as unique features and present a novel solution, working in synergy with state-of-the-art 2D object detection, to find and reconstruct in 3D, grape bunches scanned in the field by a modern smartphone. An R^2 value of 0.946 and an average precision of 0.970 was achieved when comparing our result to manual counts. Furthermore, our novel size estimation algorithm was able accurately to estimate berry sizes when manually compared to matching colour images. This work represents a novel and objective yield estimation tool that can be used on modern smartphones equipped with 3D cameras.

Occlusion of grape bunches due to foliage remains a challenge for automating grape yield estimation using computer vision. It is not always practical or possible to move or trim foliage prior to image capture. To this end, research has started investigating alternative techniques to *see* through foliage-based occlusions. This thesis introduces a novel ultrasonic-based approach that is able to volumetrically visualise grape bunches directly occluded by foliage. It is achieved through the use of a highly directional ultrasonic phased array and novel signal processing techniques to produce 3D convex hulls of foliage and grape bunches. We utilise a novel approach of agitating the foliage to enable spatial variance filtering to remove leaves and highlight specific volumes that may belong to grape bunches. This technique has wide-reaching potential, in viticulture and beyond.

Authors' Declaration

This thesis has been produced according to Massey University's "Doctoral Thesis with Publications" guidelines. Chapter 2 contains early work formatted as an article but was not published. Chapters 3, 4 & 5 consist of works that have been published in Q1 Journals. Therefore, the contents are the same as the published manuscripts but the chapters may be presented in a different style.

Acknowledgements

First and foremost I must acknowledge the unwavering support of my wife. From the compassionate encouragement to the timely "get some work done" kick up the butt. I would not have been able to make it to the end without you. You have graciously forgone precious weekends together, been understanding when I've had to work through the night, and put up with my general lack of self preservation when engrossed in a problem that I just can't let go. Keeping on track throughout the pandemic was a challenge. But you took it in your stride when research spilled into the living room and our fridge was consumed by grapes for weeks on end. I thank you deeply, for encouraging me to chase every opportunity, even when the outcome is uncertain and the path unknown.

I can't go any further without acknowledging my primary supervisor, Dr Mathew Legg. Never have I met such a patient and understanding person. Dr Legg has persevered through the complexities of research in a pandemic and the realities of life. His encouragement and support have been invaluable, and his investment in my journey goes beyond what should be expected. I am incredibly fortunate to have been able to grow as a researcher and a professional under his tutelage.

When Mathew and myself strike out on a problem, my co-supervisor, Associate Professor Fakhrul Alam is the ace up our sleeve. His door is always open to us. His excitement for research and teaching is infectious and his patience rivals that of Mathews. Dr Alam has always encouraged me to tackle new opportunities and I will always be thankful for those you afforded me and the support you gave me to succeed. You are a credit to the University and an inspiration to all you meet.

I am fortunate that my time on this journey has afforded me the pleasure of working alongside some truly amazing people.

Tyrel Glass, from contemplating the realities of life to discussing the complexities of cow behaviour, we have got up to a lot together over the years and I wouldn't trade those times for

anything.

Dr Daniel Konings, the late nights building RF data sets, and the infectious enjoyment for research and teaching. You gave me the confidence to take on this PhD and were a source of sanity when things got tough.

Dr Nathaniel Faulkner, not one thing have you put your mind to that you have not gone above and beyond with. I quickly lost track of the number of wild things you accomplished "*over the weekend*". You are a true inspiration and an absolute joy to work with.

Dr Adli Hasan Bakar, you're an inspiring testament to hard work and determination. I have thoroughly enjoyed getting to know you over the years and seeing you progress from undergrad to doctorate.

To Ruth Brooks for always being there to listen, Tanisha Pereira for the constant positive encouragement, and the many others whom I am privileged to call colleagues and friends at Massey Albany, thank you.

And finally, to my parents. Thank you for pushing me to do my best in all things, encouraging me to be forever curious, supporting my passions, and teaching me to believe in myself.

"*Look, it's just a doctorate. You don't need to thank everyone.*"

Why not?

"*Can you make it shorter then?*"

This is the short version.

Funding

This PhD has been funded by the *Rod Bonfiglioli Scholarship* provided by the *New Zealand Winegrowers Association*. This work would not have been possible without their generous support.

Table of Contents

Chapter 1

Introduction

Background

Vineyard yield estimations is an important consideration for viticulturists and the wineries they supply.An accurate forecast can reduce costs associated by allocating correct vat space, purchasing barrels and bottles, and organising correct labour requirements [1]. Also important is the resulting quality of the grapes themselves. Grapes possessing higher qualities (in such aspects as sugar content, pH level, acidity, and aroma) can produce a premium wine [2, 3, 4]. Unfortunately, it is impractical for winemakers to assess grape bunches singularly. Instead, their qualities are traditionally determined over an entire harvest 'parcel' (block of land) where the lowest common denominator of quality is often applied. Therefore, it stands to reason that viticulturists will aim to produce grapes of a uniformly high quality in order to maximize profit. However, vineyards are spatially variable. It is difficult to produce uniform crops across the entire range of soil conditions and other spatially variable influential factors [5]. "Precision Viticulture" describes a range of approaches and management tools aimed to alleviate these problems [6].

Wine grapes often provide the best quality under sub-optimal conditions. Vineyard managers will often pursue this high-quality wine using precision viticulture techniques such as deficit irrigation [7], cluster thinning [8], and variable rate applications [9]. Such techniques are highly dependent on an accurate spatial understanding of the current state of vine growth. For example, the execution of selective harvesting requires understanding the spatial distribution of the quality of grapes within the vineyard. The closer to harvest that this is performed the more informative it will be. Likewise, it is beneficial to know how environmental events, like unexpected rainfall,

may have impacted grape development. A sudden downfall can have a significant impact on deficit irrigated vines [10]. Knowing early what changes have occurred enables managers to implement corrective actions.

Traditionally, these vineyard assessments are conducted manually with trained staff. Therefore, it can be an expensive endeavour to adequately record spatial and temporal variations. As a result, the implementation of precision viticulture techniques can be seen as unprofitable by some vineyards [11, 12]. Furthermore, the subjective nature of manual assessments can add uncertainty to the results. Therefore, an automated objective and low-cost approach to yield estimation is required. We believe that computer vision and novel ultrasonic techniques provide an opportunity for such tools to be developed.

The following sections provide context to the challenges facing viticulturists, the motivations and factors at play, and what tools they currently have available.

The Variability of Vineyards

Vineyards, like most agricultural systems, are variable. For most of history, managers have simply accepted this and done the best they can on the assumption that vineyard parcels are mostly homogeneous. In 1999, the first commercial GPS harvest yield monitor was released to the market, allowing farmers to model yield as a spatial and temporally changing variable [13]. Since then, there have been numerous studies conducted into spatial and temporal variability within vineyards. In a study by Bramley et al. [14] of a vineyard in Australia, it was found that, in any given year, the yield was observed to vary by up to ten-fold. Additionally, the study highlighted significant spatial variation across the entire 7.3-hectare block. This variation, coupled with the fact that the quality of wine from these zones differed, made it an ideal candidate for zone-based vineyard management and the implementation of approaches such as selective harvesting [5].

Grape and wine quality is also highly variable and yield can not be directly linked to the quality of a wine [5, 15, 16]. Additionally, quality is not a byproduct of lower yields [14]. A study by Matthews et al. [16] concluded that the viticulture practices used to control for yield may be

more important than the yield itself as a factor for predicting grape quality. Moreover, in the study by Bramley et al. [5], it was found that between zones of similar yield, the wine was not consistent between seasons. This relationship is important, as in some situations it may be detrimental to focus on achieving high yields if it comes at the expense of quality. Therefore, by accurately monitoring vine yield throughout the season, more informed management decisions can be made to bring these factors into balance.

Given the complexity of vineyard variability, one may wonder if attempting to monitor or control for this variability is worth the investment. Indeed, Whelan and McBratney [17] define the null hypothesis of precision agriculture as "*given the large temporal variation evident in crop yield relative to the scale of a single field, then the optimal risk aversion strategy is uniform management*". However, in a study of the benefits, Bramley et al. demonstrate how the implementation of precision viticulture techniques was able to increase profit margins by $30,000 per hectare [12], a substantial return on investment.

Precision Viticulture Techniques

Precision Viticulture is a term used to describe a selection of tools and processes for optimising efficiency, increasing yields, and improving the quality of grapes within vineyards. Matese et al. [18] define precision viticulture as "*aim*[ing] *to maximize the oenological potential of vineyards*". A range of technologies and processes have been developed for this goal.

Selective harvesting is an approach used by precision viticulturists proven to provide significant economic benefits [12]. Bramley et al. define selective harvesting as "*the split picking of fruit at harvest according to different yield/quality criteria, in order to exploit the observed variation*". Traditionally this is achieved with a comprehensive survey of the crop quality performed a week or two before the planned harvest date for each parcel. The goal of this survey is to determine the spatial variation of the quality of the fruit within the parcel and create a classification map that can be then used to harvest crops into different quality categories. Typically only two categories are used for simplicity. Selective harvesting can be achieved by harvesting each categorized crop into separate bins on the same day. Alternatively, it can be

achieved by performing the harvest for each zone at different times. The latter can be used to allow lower-quality grapes to mature for slightly longer changing the quality of the final wine.

Variable rate application refers to the precise application of inputs (such as nutrients or water) to a vineyard so as to best achieve the optimal crop without being wasteful [17]. Done well, variable rate applications can reduce the spatial variability within a vineyard by targeting only the vines that need it. Success relies on a detailed understanding of the spatial and temporal variability, often to the accuracy of an individual vine. Several studies have compared the effectiveness of traditional uniform application to variable rate application as enabled by real-time modulation of spray flow rates [9, 19]. The results show significant improvement in foliage coverage and average volumetric liquid savings of up to 58% when compared to a uniform application approach [19].

Cluster thinning is a process used to reduce the crop load on a vine. The theory is that by reducing the crop load, the resources of the vine are better utilised for the development of the remaining berries [10]. Cluster thinning is traditionally done by hand, which is a laborious and time-consuming task. Vineyard managers attempt to find a balance between a reduction in yield against the potential gain in quality and additional price that might be reached. A well-timed thinning can be used to advance berry ripening for red cultivars and increase soluble solids in white varieties [8]. Effective thinning relies heavily on accurate yield estimations to develop an understanding of the current crop loading.

Computer Vision for Viticulture

In the previous section we saw that there are a number of challenges facing vineyards where automation would be beneficial. Computer vision provides potential opportunities to facilitate this automation through automated measurements. This section presents a light overview of typical computer vision approaches to automating yield estimation and the areas of active research. For a more in-depth background, the reader is directed to the literature reviews in subsequent chapters.

Grape berries grow in bunches; are relatively small in comparison to other fruits; and for coloured varieties, remain green for the majority of their growth cycle. Computer vision techniques used in vineyards typically revolve around object detection and classification in an attempt to develop an understanding of general vine growth performance. Often, metrics such as grape cluster size, shape, compactness, and grape berry size are desirable. The international OIV descriptor list [20] is a typical target for such information. The goal of machine vision in viticulture is to provide objective information regarding vine performance for the improved management of vineyards [14]. For example, early developmental stage yield maps can be used to introduce precision viticulture techniques such as variable rate application [6, 21]. Additionally, late-season spatial maps of berry quality can be used for selective harvesting [12]. Of primary importance to viticulturists is predicting the harvest weight (W_h). In a basic form, this can be expressed

$$W_h = N_b \, W_b, \tag{1.1}$$

where N_b is the number of berries and W_b is the average weight of the berries. Determining these values accurately across the entire vineyard is the primary goal of computer vision for automated yield estimation. However, there are many difficulties in estimating these in addition to other metrics that are valuable to vineyard managers. Some of the key areas of research are introduced in the following sections. Detailed reviews of applicable literature can be found in subsequent chapters.

Machine Learning and Object Detection

Early research in this space focused on the challenges that arise when applying traditional computer vision techniques to identifying grapes and grape bunches. Particular attention has been placed on colour invariant techniques due to the variable colour of grapes and the natural lighting conditions in vineyards [22, 23, 24]. Similarly, traditional algorithms for identifying circular objects, such as Hough Transform [25] and exploiting radial symmetry of grapes in controlled lighting environments have seen considerable study [26, 27, 28, 29].

The recent explosion in the capabilities of machine learning, and, in particular, object detection

algorithms has seen much of this early work rapidly superseded. Today, machine learning algorithms enjoy widespread use in industry and research focus has moved on to how best to apply these as tools.

In viticulture, machine learning techniques have traditionally been demonstrated for the detection of grape bunches in 2D colour images. Examples included convolutional neural networks [30, 31] and various YOLO (You Only Look Once) variants [32, 33, 34, 35]. For accurate yield volume estimations, it is also desirable to count the number of berries within bunches. YOLO architectures have proved effective for this purpose also [36, 37]. Research continues into using machine learning classifiers to understand the quality of individual bunches and potentially identify ailments [38, 39].

The biggest challenge with implementing these models remains the building of adequate datasets that capture enough variability of the intended target object and environment. Manual labelling is a common but very time-consuming approach and a viticulture context often needs to be repeated for different grape cultivar varieties. Techniques to automate this in a practical way promise to save a lot of time.

Measuring Berry Size

Berry size is an important indicator of the stage of growth and quality of a grape [16]. There is an inherent relationship between a berry's size and its weight. If the size of berries within a bunch can be estimated, then the harvest weight can be predicted [25, 40]. Furthermore, the distribution of berry sizes within bunches and by extension vines can be a strong indicator of vine health and expected yield and quality [41]. Grape bunch density has been deemed one of the primary descriptors of cluster growth stages by the international organization of vine and wine (OIV); presented as indicator 204 in the OIV descriptor list [20]. Understanding the growth rate of grapes between fruit set and verasian allows the implementation of precision viticulture practices such as variable rate application and cluster thinning [4, 6, 21]. However, gaining an understanding of grape-size variability within a vineyard is a laborious process, traditionally requiring the manual picking and measuring of berries and berry clusters.

This presents an excellent opportunity for computer vision-based solutions. However, for 2D approaches, resolving the absolute size of a grape requires information on its distance relative to the camera. To overcome this, some researchers modify the environment when capturing 2D images and place reference objects of known size next to the grapes. By comparing the size of the grapes to the known size of the reference object, the size of the grapes can be estimated [25]. Additionally, controlling the distance that grape bunches are positioned relative to the camera can be an effective alternative [42, 43].

Modifying the environment to estimate depth and berry size is possible with handheld solutions. However, it's not entirely practical for large-scale automated solutions. Synchronised pairs of cameras have been demonstrated as one means of solving this [40, 26]. Matching grapes can be identified within both images and using the spatial relationship between cameras the scale and physical size of individual grapes can be found.

Grape bunch and berry size can be more readily estimated through the capture of 3D information. With a sufficient quality of 3D scan, individual berry sizes can be determined with a high degree of precision [44]. Researchers have traditionally used multi-view stereo [45], and structure from motion [46] for this task due to the accuracy of 3D scan that can be achieved. However, the complex processing required to achieve these results is less than ideal. The introduction of 3D cameras presents an opportunity for efficient real-time size estimations. Often the precision is only suitable for coarse bunch size estimates [47] as the resolution to resolve individual berries is lacking at distance [48]. The quality of 3D scans of grape bunches produced by these means needs further exploration.

Portable Techniques

Numerous researchers have explored robotic solutions for automating grape yield measurements, ranging from robots capturing images by traversing vineyard rows to those autonomously picking grape bunches [26, 40, 49, 50, 47]. While these robot implementations offer advanced capabilities, they can be costly and unnecessary for many vineyards. In such cases, a more practical approach may be to utilize smartphones or handheld consumer cameras for objective, low-cost, and efficient

measurements [51, 43, 52, 53, 29]. Using a smartphone in this context has the advantage of widespread ownership among individuals, eliminating the need for growers to invest in extra equipment.

Previous studies have demonstrated complete smartphone-based solutions [27, 28, 43], but often the captured images are processed offline by more capable computers [54, 52]. Moreover, the cloud processing capabilities of modern smartphones allow flexibility for solution architects to implement more advanced algorithms.

Adopting a portable approach offers unique benefits, as it allows researchers to manipulate the grape bunch of interest. Techniques such as placing coloured cards behind the bunch to simplify segmentation or adding a known object to the scene to determine scale have been utilized [25, 43]. Additionally, using an artificial background held at a fixed distance from the camera enables consistent scaling of grape bunches in relation to the background and thus grape size can be inferred [42].

3D Systems for Precision Viticulture

Applying traditional 2D computer vision techniques to precision viticulture continues to be a large focus for research. In particular, the application of the latest machine learning tools continues to unlock new potential. However, 2D solutions lack information about the scale of the environment and continue to rely on optimal lighting and environmental conditions to achieve good results. Recent advancements in 3D sensors and the introduction of low-cost real-time 3D cameras have opened the field to new techniques that will lead to more robust approaches and increased accuracy.

Historically, the use of 3D information in vineyard canopy measurement has played a valuable role in optimizing crop loads and implementing variable rate applications to enhance efficiency [55]. Notably, there have been successful applications of low-cost ultrasonic transducers [56, 57] and push-broom style LiDAR scanners [58] for this purpose. These technologies have demonstrated significant volume savings of up to 58% by adjusting the spray flow rate of fertilizers

[19, 9]. Unlike ultrasound, the accuracy and precision of LiDAR scanners make them useful beyond canopy estimation. Their use has been explored for estimation of 3D biomass of wintered vines [59], canopy fill characteristics, and leaf wall area [55]. However, effective use requires significant post-processing and careful alignment of the scan data [56, 59].

On a larger scale, 3D information has been used as a tool for surveying entire vineyards using aerial Structure-from-Motion (SfM) techniques and low-cost consumer drones. This approach has demonstrated the collection of numerous critical metrics for vineyard management including canopy volume, vine height, vine width, and growth uniformity [60, 61]. However, its success is highly dependent on image quality [62]. SfM has also been combined with hyperspectral imaging [63] to model leaf area index (LAI) comparable with ground-based approaches and construct hyperspectral digital surface models (HS DSM) [64, 65].

SfM has also seen the use on ground robots to capture detailed 3D point clouds of entire vineyard parcels [45, 46]. The resulting point clouds are then processed further to segment different aspects of the vines' structure including branches, leaves, and grapes [66, 67]. However, the process of generating these point clouds takes time which limits its use to post-capture analysis [68] and restricts its use in real-time robotics or in situ precision viticulture applications.

For grape yield estimation, a key benefit 3D information provides over traditional 2D imaging techniques is the potential for the size of grapes and bunches to be determined. This gives rise to useful metrics including bunch compactness, average berry size, bunch volume, and berry weight. Furthermore, capturing the surface curvature of grapes enables shape-fitting algorithms such as Random Sample and Consensus (RANSAC) to be applied for individual berry identification [69, 70]. Depth information used in conjunction with traditional 2D imaging also has merit, for example, the automated removal of backgrounds based on depth thresholds to simplify 2D image processing [71].

Light Detection and Ranging (LiDAR) based solutions have also been explored for use in viticulture. These typically differ from the push-broom style scanners discussed above in that they capture 3D information from a perspective akin to a traditional camera. Highly accurate scanners have been used in lab conditions to capture complete 3D scans of grape bunches for

high-precision phenotyping [69, 70]. The accuracy of these scans enables direct identification of individual grapes through RANSAC sphere [72] and ellipsoid [73] fitting. Furthermore, full reconstruction of the internal bunch structure has been demonstrated using morphological growth models and restricted reconstruction grammars [74]. The laser scanners used in these studies are expensive and require skilled manual operation. They have been tested in field conditions, but their sensitivity to the movement of the grape bunch and surrounding foliage complicates the process.

Over the last few years, there has been a significant rise in the availability of low-cost 3D cameras. These cameras offer a distinct advantage by capturing real-time 3D information about a scene. This makes them uniquely suitable for robotics and other time-critical applications [75] but also as a more efficient means of building large 3D scans for post-processing [76]. The real-time nature of 3D cameras lends itself to robotic applications such as bunch pose estimation [77] for autonomous vine pruning [75]. Additionally, their application towards characterising vine canopies [78] for variable rate applications has been explored. Several 3D camera technologies have been developed and their performance within a viticulture context has been explored to varying degrees. Time of Flight cameras have received particular interest due to their superior performance in difficult lighting conditions [79]. However, noise and accuracy remain a challenge. Other 3D camera technologies have been featured in literature, including Structured Light (SL) [48, 80] and Active Stereo Vision (ASV) [47, 81]. However, these struggle in outdoor environments with direct sunlight due to saturation of their emitted IR patterns [82]. Cost-effective 3D cameras that utilise LiDAR principles have recently become available. Initial studies have shown performance comparable to ToF cameras; however, direct sunlight presents a challenge in some cases [79].

Addressing Occlusion

A limitation of computer vision techniques is that it requires the complete visibility of grape bunches for accurate results. However, this is not reliably achievable due to occlusions present in vineyard. Therefore, occlusion may be the biggest unsolved issue when it comes to automated solutions for yield estimation.

According to the OIV descriptor list, a grape cluster's size, compactness, and shape are vital to understanding the vine's growth and performance [20]. Thus, understanding the structure of each entire bunch is crucial for effective analysis. Occlusions take multiple forms in the context of viticulture. Nuske et al. [83] present three occlusion types that are the source of the majority of errors: self-occlusion, bunch-occlusions, and foliage-occlusions.

- The complex nature of grape bunches results in self-occlusions where the visible part of the bunch may not be an accurate representation of the whole.
- Grape bunches often grow in close proximity to other bunches. This often leads to bunches occluding and, in some cases, being indistinguishable from each other from a single perspective.
- Grape bunches partially or fully occluded by foliage present a challenge for machine vision systems attempting to present accurate yield estimates.

Self occlusions Grape bunches can host upwards of 100 individual grapes and grow in a large variety of shapes depending on variety and environmental influences [41, 20]. As with general foliage occlusion discussed prior, methods of dealing with self-occlusion typically also involve a visibility factor (V_f) [84]. This naively assumes the visible portion (N_b^v) of the bunch is proportional to the actual berry count (N_b) using

$$N_b = V_f \cdot N_b^v. \tag{1.2}$$

However, it is understood that such a factor attempts to account for too many sources of variation [83]. Alternative approaches, such as modelling the bunch as a 3D ellipsoid with a size proportional to the number of berries within the bunch have been demonstrated to be more accurate [83]. However, these require well-defined bunch dimensions that are difficult to obtain from 2D images. Additionally, the application of morphological growth models to the visible portion of grape bunches shows great promise for estimating the structure of the non-visible portion. To date, such approaches have utilised manual measurements taken in controlled conditions [85, 86] or complete 3D laser scans of grape bunches [74].

Foliage occlusions To reduce the impact of foliage occlusion, images will often be taken after manual removal of foliage around the grapes [87, 66]. However, this can impact the crop-loading potential of the vines and may not be a possible solution at some stages of vine growth or may not be desirable for specific species. Furthermore, manual involvement presents a bottleneck for autonomous solutions.

In most computer vision solutions, foliage occlusions are accounted for by a simple scaling "visibility" factor, deduced from manual measurements [40]. However, this approach is naive and its accuracy will likely vary throughout a vineyard. Furthermore, a simple scaling factor only attempts to account for yield variations. Important metrics such as the size and shape of a cluster will remain impacted. More advanced methods have been demonstrated that model the expected proportion of occluded bunches by analysing the canopy porosity and visible bunch area [88]. Such an approach has the ability to adapt its visibility factor to spatial variability across the vineyard but needs to be tuned to different cultivars and stages of growth.

Recently, researchers have attempted to see through foliage occlusions with the development of microwave radar-based yield prediction [89, 90]. However, these technologies are currently costly, and still far from a commercial solution. In a similar fashion, studies have demonstrated the potential for ultrasonic phased arrays to image through pasture [91]. However, no previous work has explored such a technique for detecting grapes occluded by foliage. Integration of ultrasound with traditional computer vision techniques may facilitate targeted ultrasonic scans or classification of ultrasonically scanned volumes.

Motivation

Numerous computer vision algorithms have been developed for the detection of grape berries within images. Approaches vary from using colour segmentation and edge detection to training object detectors such as the ever-popular YOLO convolutional neural network models. The state of the art in these approaches have been shown to accurately identify over 90% of visible berries. However, even with these results, yield estimations are still prone to errors. One of the primary reasons is that the majority of berries are occluded from the camera's view, often

due to foliage and other obstacles. In such situations, it is common for researchers to apply a correction factor that approximates the proportion of occlusion over the entire vineyard. This is less than ideal if there is a spatial variation of this factor throughout the vineyard. Complexity is also introduced when finding the size of individual grapes is desired. This can be useful for mapping growth rates and identifying the distribution of Hen and Chicken (Millerandage) in bunches. In these cases, techniques for controlling the backgrounds behind bunches of interest and adding calibration objects of known size to the scene are common [25, 42].

Yield Estimation with 3D Cameras

These limitations have initiated a body of research to explore the use of 3D information for improving berry identification. Utilising 3D information, researchers can apply 3D shape-based feature detection to address colour and illumination issues. Additionally, by combining multiple viewpoints into one 3D reconstruction, small occlusions can often be removed. However, one of the largest benefits of utilising 3D information is that berry size can be readily established without needing calibration objects in each image. The addition of 3D information in this way has demonstrated improved results over traditional 2D methods [87].

Structure-from-Motion (SfM) is a technique that captures 3D information derived from a sequence of RGB images. These images are processed to find correspondence between image pairs and create dense point clouds. However, such approaches are computationally expensive and, for a single row of vines, may take hours to compute [45]. An alternative approach is to utilise 3D cameras which are able to produce 3D images of a scene in real-time. This is altogether faster and more flexible, opening the door for integration with real-time robotics and active vision techniques [47, 77, 92]. Such 3D cameras are relatively new and not much work has been done towards their use in agriculture.

This work aims to investigate the feasibility of using, low-cost 3D cameras for the purpose of yield estimation. Compared to current approaches within the literature, these have the advantage of being highly portable and working in real-time. To date, there has been little investigation into the use of these cameras for vineyard yield estimation. We believe there is significant potential

to leverage the strengths of these 3D cameras for objective mobile assessments by utilising 3D cameras within modern smartphones as well as exploiting their real-time nature for robotic active vision systems.

Ultrasonic Detection of Occluded Grapes

Addressing foliage occlusions poses a significant challenge for automation using conventional computer vision or 3D cameras. In many situations a technology that enables seeing through occlusions and accurately identifying grape bunches would be highly desirable. A promising solution to tackle this issue lies in employing ultrasonic phased arrays. These arrays possess a narrow beamwidth, which allows them to distinguish between multiple echoes effectively. Their successful application in pasture bio-mass estimation, where they can image through foliage to the underlying ground layer, demonstrates their potential.

However, despite the success in pasture applications, no previous studies have explored the use of ultrasonic phased arrays for detecting occluded fruit, such as grape bunches. This research aims to investigate the feasibility of employing this technique for that purpose and proposes innovative methods to filter out foliage occlusions, enabling the identification of occluded grape bunches.

Research Questions

This research aims to address the following questions:

- How do low-cost 3D camera technologies perform for imaging individual and clusters of spherical objects?
- How do low-cost 3D camera technologies perform when imaging grape bunches in lab and outdoor environments?
- Can scans from 3D cameras effectively identify and estimate the size of individual berries in unstructured and uncontrolled field conditions?
- Can air-coupled ultrasonic phased arrays identify grape bunches occluded by foliage?

Scope of the Study

Grape yield estimation is a complex process that is applied throughout the world with various different approaches and techniques. Ultimately, the goal is to increase the value of the harvest and optimise the logistics surrounding it. Typically this is a manual process and relies on the intuition and understanding of skilled viticulturists. Standardisation, regularity and vineyard coverage are all areas where automation is being looked to for solutions. Yield estimation is a process performed regularly throughout the growing season. From initial bud breakthrough to days before harvest. Therefore, different algorithms are needed at different periods within the season. An algorithm for detecting and classifying the health and distribution of buds will be different from one detecting growth and distribution of berries pre-veraison, and that again is likely to be different from mid and post veraison where the colour and translucency change significantly.

In a similar way, the grape varieties being inspected may require changes to techniques and approaches. Most table grapes are grown for their large size, but grapes grown for wine often favour smaller berries with strong flavours.

Vine management is also a key consideration, different varieties of grapes are managed differently even though the same varieties may have different practices in different parts of the world. A simple example is differing levels of foliage cover maintained throughout the season. This directly impacts the proportion of visible and occluded grape bunches. Alternatively, the type of trellis used will impact the position of grapes within the vine and again may result in occlusions or other complicating factors at play.

Finally, the seasonal nature of grape growing impacts the research windows available. Understanding the growth rate of grapes between fruit set and veraison allows the implementation of precision viticulture practices such as variable rate application and bunch thinning [4, 6, 21]. Understanding the maturity and growth rates through veraison up to harvest helps with selective harvesting practices [12]. Capturing datasets of grapes in the post veraison leads up to harvest is a narrow window of time occurring once a year. Furthermore, once samples are taken there

is a very small window of time that they can be analysed in the lab. Fortunately, post-veraison grapes are very similar to table grapes (aside from their size), so in many cases, grapes available at supermarkets are suitable analogs. Capturing data sets of vine buds or berries pre-veraison presents a more difficult situation there is no suitable year-round analog that can be utilised.

In this study, we have captured data sets of grapes captured in field conditions in the post veraison stage up to three weeks before harvest. This was done due to the timing of the research, we were ready to capture data in March 2020 just before harvest. Unfortunately, then COVID hindered subsequent vineyard visits for later in the year. For the remainder of the studies, table grapes were used as accurate proxies for the smaller wine grapes. No work was done on bud or berries pre-veraison due to the inability to collect data sets. Therefore the scope of this work focuses on berries post-veraison, both pre & and post-harvest. Wherever applicable, the differences between table and wine grapes sampled from the field have been commented on. The datasets collected of wine grapes in field conditions during the windows available are extensive and have not yet been fully explored in this work. We expect they will provide a solid foundation for future work in this space.

Publication List

[93] B. Parr, M. Legg, F. Alam. Analysis of depth cameras for proximal sensing of grapes. Sensors (2022) 22(11), 4179. **doi:10.3390/s22114179**

[94] B. Parr, M. Legg, F. Alam. Grape Yield Estimation with a Smartphone's Colour and Depth Cameras using Machine Learning and Computer Vision Techniques. Computers and Electronics in Agriculture (2023) 213, 108174. **doi:10.1016/j.compag.2023.108174**

[95] B. Parr, M. Legg, S. Bradley, F. Alam. Occluded grape cluster detection and vine canopy visualisation using an ultrasonic phased array. Sensors (2021) 21(6), 2182. **doi:10.3390/s21062182**

[96] B. Parr, M. Legg, F. Alam. Analysis of depth cameras for proximal sensing of grapes. IEEE Sensors Applications Symposium (2020) **doi:10.1109/SAS48726.2020.9220078**

Thesis Overview

Chapter 2

Chapter two addresses our first research question by presenting an exploratory analysis of the performance of a range of popular 3D cameras for the purpose of capturing spherical objects in a controlled environment. The work compares different 3D camera technologies including passive and active stereo, structured light, and time of flight. Previous camera performance studies identified in literature have primarily focused on the accuracy and precision of correctly capturing planer surfaces perpendicular to the camera's view [97]. However, these do not present an accurate analog for what performance may be expected when capturing spherical surfaces, as is the case for grapes. In such a situation, accurate depth estimates are required over the entire surface to correctly interpret the size and shape of the spherical object. Therefore, this work looks to identify and measure the fundamental performance metrics required to accurately capture spherical surfaces and lay the foundation for our understanding of the best-case performance of each technology.

This work has the following contributions:

- **Comparison of projected depth errors with respect to the incidence angle.** This work presents the magnitude and distribution of depth errors with respect to the normal angle of the surface that is being measured at each point. The more of the grape's surface curvature that can be accurately detected, the more information will be available to be used in subsequent algorithms. Cameras with poor performance at high incidence angles are likely to present surfaces with a larger apparent curvature than really exists. This is especially vital for larger measurement distances or low-resolution cameras where the point density is relatively low.
- **Comparison of projected depth errors with respect to the surface convergence.** Grape bunches are complex 3D objects that can be approximated by the dense packing of 3D spheres. Therefore it is vital that 3D cameras are capable of accurately measuring the converging surfaces between individual berries. This work presents a new approach to

measuring accuracy in these situations by computing a local convergence metric for each point. We then present a comparison of the ideal versus measured convergence for each camera at a range of distances.

- **Analysis of flying pixels.** This work presents an analysis of erroneous points; points incorrectly identified as belonging to the reference surface. These are often called flying pixels as they manifest in regions of depth discontinuity such as on the edges of objects. We present these in the form of a Hit to miss ratio to normalise for the pixel density of each camera.

Chapter 3

The results presented in Chapter 2 provide confidence that 3D cameras have the capability to capture the surface of spherical objects albeit with compromises depending on the technology chosen. In Chapter 3, we address our second research question by extending this work to bunches of grapes and a wider range of 3D cameras in both indoor controlled and outdoor environments. We explore the impact of the translucent surface of grapes and highlight the distortions present in the time of flight-based technologies due to subsurface scattering. We then assess each technology's merits for individual grape detection using a traditional sphere-fitting approach.

This work was published in a Q1 Open Access Journal, MDPI Sensors, and has the following contributions:

- **The first study to benchmark the performance of multiple RGB-D cameras for use in grape yield estimation applications.** This study is the first to compare the performance of multiple RGB-D cameras in grape yield estimation applications, including ToF cameras that have not been previously used in such studies. The performance analysis involved calculating error maps by comparing high-resolution scans obtained through photogrammetry with those obtained through the RGB-D cameras. The evaluation also considered the cameras' performance under both direct sunlight and shaded conditions.
- **The first study to analyse depth map errors in the RGB-D scans of grapes**

at an individual grape berry scale. The ability to distinguish individual grapes from 3D scans offers valuable insights for estimating yield and crop load. This information can help viticulturists assess metrics like the distribution of berry sizes and the number of berries per bunch. Additionally, utilizing 3D modelling to analyze the architecture of grape bunches holds the potential for more precise volume estimations.

- **The first comparison of the ability for RGB-D cameras to detect individual grape berries using Random Sample Consensus (RANSAC).** RANSAC is a well-understood method for fitting objects to point clouds [70]. However, this study is the first to directly compare its performance across different low-cost 3D cameras that utilize a range of different technologies.

- **The first presentation of the distortion of grape berries in ToF and LiDAR scans.** Time-of-Flight (ToF) and LiDAR 3D cameras measure distance by reflecting light off the surface of objects. In this study, it has been observed for the first time that 3D cameras using these technologies can result in distorted representations of the grape surface when light enters the grape berries. The translucent properties of grapes mean that diffused scattering occurs within the berries, which causes distortion of the ToF and LIDAR 3D camera results. Interestingly, the nature of the distortion appears to vary depending on whether all pixels are measured simultaneously (ToF) or independently (LiDAR). These distortion profiles suggest that they could be utilized as a potential means of non-destructive testing to understand the characteristics of each grape berry.

Chapter 4

Chapter 3 highlighted for the first time that subsurface scattering within the berries causes distortions that manifest as peaks in time of flight based 3D scans of grapes. In Chapter 4, we exploit these distortions to address our third research question and develop a mobile application for grape identification and size estimation using 3D and RGB cameras present in a modern consumer smartphone. Peaks identified in 3D scans captured of grape bunches in field conditions are compared to manual counts. We then utilise these peaks in captures of grape bunches in lab

conditions to build an unsupervised YOLO v7 model capable of accurately identifying individual berries within RGB images captured by a smartphone. By then finding corresponding peaks in the depth image we demonstrate a unique approach to estimate berry sizes and generate 3D reconstructions of the visible grape bunch.

This article has been published in a Q1 Open Access Journal, Computers and Electronics in Agriculture, and has the following contributions:

- **The first application of a built-in depth camera within a smartphone for grape yield estimation.** This work presents the first time a depth camera within a modern smartphone has been used in field conditions for objective grape yield estimation and bunch analysis. Previous works have only made use of RGB cameras and typically require structured environments with controlled backgrounds [43]. This approach presents the state-of-the-art as 3D cameras in mobile phones start to become ubiquitous.
- **Novel technique to identify individual grape berries in the ToF depth maps using distortions caused by diffused scattering.** Expanding on our previous discovery of peak-like distortions in ToF scans of grapes. We present a novel technique to identify grapes using a simple peak detection approach coupled with intuitive spatial filtering techniques. A precision of 89% was achieved in tests performed on unstructured field images. The results suggest that occluded berries on the edges of bunches are likely to be missed when solely relying on detecting prominent peaks. Future work will explore methods of detection that are more robust in these situations.
- **The first unsupervised training of a YOLO model for detecting grape berries using 3D-based labels.** Typically object detection models are trained on manually labelled data. This can be a time-consuming process. In this work we exploit the grapes detected in depth images to automatically build a YOLO V7 training set for corresponding RGB images. We utilise an unsupervised autoencoder as a pre-label filter to remove obvious outliers and utilise depth thresholding to mask out the bunches in the RGB image to substitute different backgrounds and build a model robust to real-world environments. This technique produced a model with a precision of 97% when tested on images captured

in uncontrolled & unstructured field environments despite not being trained on any such images. The highly automated nature of this approach means it can be easily extended to different cultivars or even adapted to different crops with ease.

- **A novel modelling technique to estimate the size of berries in a bunch using the measured 3D locations of the berries found using peak detection.** A new 3d grape bunch modelling technique is presented that estimates the size of berries in a bunch using the measured 3D locations of the berries. Given an approximate starting size for all berries in the bunch and each berry's 3D peak, it is able to quickly and stably converge on an estimate of each berry's size. In empirical comparisons, this approach appears to work well even in bunches with a significant variance in berry sizes. Initial size estimates for each berry can be supplied to improve convergence speed but these are not needed and hence this approach and be run independently of RGB-based size estimates.

Chapter 5

In Chapter 4, we presented an effective approach to identifying and assessing grapes in field conditions for the purpose of objective yield estimation with a standard smartphone. In such a situation, the user is able to manually ensure an unoccluded view of the grape bunch. However, if we were to extend 3D scanning to an autonomous platform, this process would be difficult to automate and a clear view of grapes is not guaranteed. Utilising 3D reconstructions of what is visible presents interesting options for bunch reconstruction [98] or robotic active vision techniques [92] but this will require further study. Chapter 5 address our fourth research question and presents a novel alternative approach to the problem of occlusion through the use of an ultrasonic phased array with beamforming techniques and probabilistic volume thresholding. With this approach, we demonstrate the ability to see through foliage occlusions and identify volumes of potential grape bunches. It is feasible to imagine such an approach being used in conjunction with 3D cameras and active vision techniques to provide an effective large-scale automated yield estimation process that is robust to occlusions.

This work was published in a Q1 Open Access Journal, MDPI Sensors, and has the following

contributions:

- **The first work to have used an air-coupled ultrasonic phased array and coded waveforms for the purpose of analysing vine canopies.** Ultrasound ranging has been used in the past for estimating vine canopies [56]. However, these are limited to capturing a coarse estimation of the volume envelope of the foliage due to the large field of view and lack of selectivity of the transducers used. This work uses a phased array of transducers and coded waveforms to drastically reduce the beamwidth of the ultrasound signal and increase the selectivity to a level where individual leaves and other objects are distinguishable. This is the first time such a narrow beamwidth ultrasound array has been used on vines and presents exciting potential for future work.

- **The first study to investigate if it is possible to use ultrasound to image through leaves, to detect fruit located behind leaves.** Due to the narrow beamwidth of the ultrasound array we present for the first time the ability to distinguish depth between multiple near-field echos. We exploit this capability to identify objects that are visually occluded by foliage and present a case where this can be used to identify occluded grape bunches.

- **The first work to differentiate echoes that come from leaves through agitation.** Receiving multiple echos lets us build a comprehensive understanding of 3D volume of a vine beyond just the envelope. However, differentiating between surfaces is not possible without further analysis. This work present a unique technique for differentiating between echos from light objects such as leaves and those more ridged such as grape bunches and branches. This is achieved through non-destructive agitation of the foliage using moving air and then analysing the variance in the measured 3D volumes. With this technique, we show that 3D volumes relating to foliage can be removed from acoustic scans leaving behind occluded grape bunches.

- **A novel technique for improving the resolution of array-based cross-correlation for near-field echoes.** This approach simulates the effect of focusing the transmission of the array at any desired depth in post-processing. This eliminates the need for the

complex electronics required for focusing the array's transmission to a desired scan depth and means that multiple depths can be focused on with only a single transmission. This technique is the first step towards real-time detailed 3D volumetric scanning of vines and has significant implications beyond just viticulture.

Chapter 6

Chapter 6 concludes the thesis by summarising the main contributions and suggests directions for future work.

Appendix 1

Appendix 1 contains a conference paper published in the 2020 IEEE Sensors Applications Symposium (SAS) which presents our work with an ultrasonic phased array to visualise vine foliage and isolate grape bunches. It was extended into the article presented in Chapter 5 with the addition of novel beamforming distortion compensation that increased the achievable volumetric resolution.

A Note on Structure

Chapter 2 is a draft of a paper that has not been published. It has been included on the basis that we believe it to be novel work and has contributed significantly to the research journey of this doctorate. Chapters 3, 4 & 5 are peer-reviewed journal articles that have been published in Q1 Journals. Each chapter contains a self-contained literature review focused on that articles contributions and specific area of research. As such, there is expected to be some repetition in the introductions and conclusions. The order of chapters is not chronological according to when corresponding work was performed but has been chosen to provide narrative structure. In particular, Chapter 5 was published before Chapter 3 due to the availability of the ultrasonic phased array.

References

[1] C. M. Lopes, J. Graça, J. Sastre, M. Reyes, R. Guzmán, R. Braga, A. Monteiro, P. A. Pinto, Vineyard yeld estimation by VINBOT robot-preliminary results with the white variety viosinho, in: Proceedings 11th Int. Terroir Congress. Jones, G. and Doran, N.(eds.), pp. 458-463. Southern Oregon University, Ashland, USA., Jones, G.; Doran, N.(eds.), 2016.

[2] R. Weaver, M. Amerine, A. Winkler, Preliminary report on effect of level of crop on development of color in certain red wine grapes, American Journal of Enology and Viticulture 8 (4) (1957) 157–166.

[3] B. Bravdo, Y. Hepner, C. Loinger, S. Cohen, H. Tabacman, Effect of crop level on growth, yield and wine quality of a high yielding Carignane vineyard, American Journal of Enology and Viticulture 35 (4) (1984) 247–252.

[4] R. Smart, J. K. Dick, I. M. Gravett, B. Fisher, Canopy management to improve grape yield and wine quality-principles and practices, South African Journal of Enology and Viticulture 11 (1) (1990) 3–17.

[5] R. Bramley, Understanding variability in winegrape production systems, Australian Journal of Grape and Wine Research 11 (1) (2005) 33–42.

[6] J. Arnó Satorra, J. A. Martinez Casasnovas, M. Ribes Dasi, J. R. Rosell Polo, Review. Precision viticulture. Research topics, challenges and opportunities in site-specific vineyard management, Spanish Journal of Agricultural Research, 2009, vol. 7, núm. 4, p. 779-790 (2009).

[7] M. Keller, Deficit irrigation and vine mineral nutrition, American Journal of Enology and Viticulture 56 (3) (2005) 267–283.

[8] T. Preszler, T. M. Schmit, J. E. V. Heuvel, A model to establish economically sustainable cluster-thinning practices, American Journal of Enology and Viticulture 61 (1) (2010) 140–146.

[9] J. Llorens, E. Gil, J. Llop, A. Escola, Variable rate dosing in precision viticulture: Use of electronic devices to improve application efficiency, Crop protection 29 (3) (2010) 239–248.

[10] M. Keller, L. J. Mills, R. L. Wample, S. E. Spayd, Cluster thinning effects on three deficit-irrigated vitis vinifera cultivars, American Journal of Enology and Viticulture 56 (2) (2005) 91–103.

[11] R. Bramley, B. Pearse, P. Chamberlain, Being profitable precisely-a case study of precision viticulture from Margaret River, Australian and New Zealand Grapegrower and Winemaker (2003) 84–87.

[12] R. Bramley, A. Proffitt, C. Hinze, B. Pearse, R. Hamilton, Generating benefits from precision viticulture through selective harvesting, Precision agriculture 5 (2005) 891–898.

[13] R. Bramley, A. Proffitt, Managing variability in viticultural production, Grapegrower and Winemaker 427 (1999) 11–16.

[14] R. Bramley, R. Hamilton, Understanding variability in winegrape production systems, Australian Journal of Grape and Wine Research 10 (1) (2004) 32–45.

[15] R. Bramley, J. Ouzman, P. Boss, Variation in vine vigour, grape yield and vineyard soils and topography as indicators of variation in the chemical composition of grapes, wine and wine sensory attributes, Australian Journal of Grape and Wine Research 17 (2) (2011) 217–229.

[16] M. Matthews, V. Nuzzo, Berry size and yield paradigms on grapes and wines quality, Acta Horticulturae 754 (2007) 423.

[17] B. Whelan, A. McBratney, The "null hypothesis" of precision agriculture management, Precision Agriculture 2 (3) (2000) 265–279.

[18] A. Matese, S. F. Di Gennaro, Technology in precision viticulture: A state of the art review, International Journal of Wine Research 7 (2015) 69–81.

[19] E. Gil, A. Escola, J. Rosell, S. Planas, L. Val, Variable rate application of plant protection products in vineyard using ultrasonic sensors, Crop Protection 26 (8) (2007) 1287–1297.

[20] O. I. de la Vigne et du Vin, Oiv descriptor list for grape varieties and vitis species (2007).

[21] A. Tagarakis, V. Liakos, S. Fountas, S. Koundouras, K. Aggelopoulou, T. Gemtos, Management zones delineation using fuzzy clustering techniques in vines (2011).

[22] G. M. Dunn, S. R. Martin, Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest, Australian Journal of Grape and Wine Research 10 (3) (2004) 196–198.

[23] M.-P. Diago, C. Correa, B. Millán, P. Barreiro, C. Valero, J. Tardaguila, Grapevine yield and leaf area estimation using supervised classification methodology on RGB images taken under field conditions., Sensors 12 (12) (2012) 16988–7006. doi:10.3390/s121216988.

[24] T. Kawaguchi, The image parsing system for grape picking robots utilizing RGBD images and superpixels (2015).

[25] E. A. Murillo-Bracamontes, M. E. Martinez-Rosas, M. M. Miranda-Velasco, H. L. Martinez-Reyes, J. R. Martinez-Sandoval, H. Cervantes-de Avila, Implementation of Hough transform for fruit image segmentation, Procedia Engineering 35 (2012) 230–239.

[26] Z. S. Pothen, S. Nuske, Texture-based fruit detection via images using the smooth patterns on the fruit, in: Robotics and Automation (ICRA), 2016 IEEE International Conference on, IEEE, 2016, pp. 5171–5176.

[27] M. Grossêtete, Y. Berthoumieu, J.-P. Da Costa, C. Germain, O. Lavialle, G. Grenier, A new approach on early estimation of vineyard yield: Site specific counting of berries by using a smartphone, in: European Conference on Precision Agriculture, 2011, pp. 8–pages.

[28] M. Grossetete, Y. Berthoumieu, J.-P. Da Costa, C. Germain, O. Lavialle, G. Grenier, et al., Early estimation of vineyard yield: Site specific counting of berries by using a smartphone, in: International Conference of Agricultural Engineering-CIGR-AgEng, 2012.

[29] A. Aquino, I. Barrio, M.-P. Diago, B. Millan, J. Tardaguila, vitisberry: An Android-smartphone application to early evaluate the number of grapevine berries by means of image analysis, Computers and Electronics in Agriculture 148 (2018) 19–28.

[30] T. T. Santos, L. L. de Souza, A. A. dos Santos, S. Avila, Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association, Computers and Electronics in Agriculture 170 (2020) 105247.

[31] L. Coviello, M. Cristoforetti, G. Jurman, C. Furlanello, GBCNet: In-field grape berries counting for yield estimation by dilated CNNs, Applied Sciences 10 (14) (2020) 4870.

[32] H. Li, C. Li, G. Li, L. Chen, A real-time table grape detection method based on improved YOLOv4-tiny network in complex background, Biosystems Engineering 212 (2021) 347–359.

[33] R. Zhao, Y. Zhu, Y. Li, An end-to-end lightweight model for grape and picking point simultaneous detection, Biosystems Engineering 223 (2022) 174–188.

[34] B. Liu, L. Luo, J. Wang, Q. Lu, H. Wei, Y. Zhang, W. Zhu, An improved lightweight network based on deep learning for grape recognition in unstructured environments, Information Processing in Agriculture (2023).

[35] L. Shen, J. Su, R. He, L. Song, R. Huang, Y. Fang, Y. Song, B. Su, Real-time tracking and counting of grape clusters in the field based on channel pruning with YOLOv5s, Computers and Electronics in Agriculture 206 (2023) 107662. doi:https://doi.org/10.1016/j.compag.2023.107662.

[36] Y. Miao, L. Huang, S. Zhang, A two-step phenotypic parameter measurement strategy for overlapped grapes under different light conditions, Sensors 21 (13) (2021) 4532.

[37] Harry, Berry YOLO V5 Dataset, visited on 2023-03-20 (Nov 2021).
URL https://universe.roboflow.com/new-workspace-hzmvk/berry_yolov5-slwnw

[38] D. Wang, W. Cao, F. Zhang, Z. Li, S. Xu, X. Wu, A review of deep learning in multiscale agricultural sensing, Remote Sensing 14 (3) (2022). doi:10.3390/rs14030559.

[39] S. M. Javidan, A. Banakar, K. A. Vakilian, Y. Ampatzidis, Diagnosis of grape leaf diseases using automatic k-means clustering and machine learning, Smart Agricultural Technology 3 (2023) 100081. doi:https://doi.org/10.1016/j.atech.2022.100081.

[40] O. Mirbod, L. Yoder, S. Nuske, Automated measurement of berry size in images, IFAC-PapersOnLine 49 (16) (2016) 79–84.

[41] J. Booysen, C. Orffer, E. Beukman, Crop forecasting for vine grapes in South-Africa, OVRI Stellenbosch, 78 94 (1978).

[42] G. Rabatel, C. Guizard, Grape berry calibration by computer vision using elliptical model fitting, in: European Conference on Precision Agriculture, Vol. 6, 2007, pp. 581–587.

[43] S. Liu, X. Zeng, M. Whitty, 3DBunch: A novel iOS-smartphone application to evaluate the number of grape berries per bunch using image analysis techniques, IEEE Access 8 (2020) 114663–114674.

[44] A. Kicherer, K. Herzog, M. Pflanz, M. Wieland, P. Rüger, S. Kecke, H. Kuhlmann, R. Töpfer, An automated field phenotyping pipeline for application in grapevine research, Sensors 15 (3) (2015) 4823–4836. doi:10.3390/s150304823.

[45] J. C. Rose, A. Kicherer, M. Wieland, L. Klingbeil, R. Töpfer, H. Kuhlmann, Towards automated large-scale 3D phenotyping of vineyards under field conditions, Sensors 16 (12) (2016). doi:10.3390/s16122136.

[46] T. T. Santos, L. H. Bassoi, H. Oldoni, R. L. Martins, Automatic grape bunch detection in vineyards based on affordable 3D phenotyping using a consumer webcam., in: Embrapa Informática Agropecuária-Artigo em anais de congresso (ALICE), Congresso Brasileiro de Agroinformatica, 2017.

[47] P. Kurtser, O. Ringdahl, N. Rotstein, R. Berenstein, Y. Edan, In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera, IEEE Robotics and Automation Letters 5 (2) (2020) 2031–2038. doi:10.1109/LRA.2020.2970654.

[48] F. Marinello, A. Pezzuolo, D. Cillis, L. Sartori, Kinect 3D reconstruction for quantification of grape bunches volume and mass, Engineering for Rural Development 15 (2016) 876–881.

[49] T. Botterill, S. Paulin, R. Green, S. Williams, J. Lin, V. Saxton, S. Mills, X. Chen, S. Corbett-Davies, A robot system for pruning grape vines, Journal of Field Robotics 34 (6) (2017) 1100–1122.

[50] A. Kicherer, K. Herzog, N. Bendel, H.-C. Klück, A. Backhaus, M. Wieland, J. Rose, L. Klingbeil, T. Läbe, C. Hohl, et al., Phenoliner: A new field phenotyping platform for grapevine research, Sensors 17 (7) (2017) 1625.

[51] J. Tardaguila, M. Stoll, S. Gutiérrez, T. Proffitt, M. P. Diago, Smart applications and digital technologies in viticulture: A review, Smart Agricultural Technology 1 (2021) 100005.

[52] S. Liu, X. Zeng, M. Whitty, A vision-based robust grape berry counting algorithm for fast calibration-free bunch weight estimation in the field, Computers and Electronics in Agriculture 173 (2020) 105360.

[53] L. Schmidtke, Developing a phone-based imaging tool to inform on fruit volume and potential optimal harvest time, Tech. Rep. Project No. CSU 1501, National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, New South Wales, Australia (June 2018).

[54] M. P. Diago, A. Sanz-Garcia, B. Millan, J. Blasco, J. Tardaguila, Assessment of flower number per inflorescence in grapevine by image analysis under field conditions, Journal of the Science of Food and Agriculture 94 (10) (2014) 1981–1987.

[55] T. Bates, B. Grochalsky, S. Nuske, Automating measurements of canopy and fruit to map crop load in commercial vineyards, Res Focus: Cornell Viticulture Enology 4 (2011) 1–6.

[56] J. Llorens, E. Gil, J. Llop, et al., Ultrasonic and LIDAR sensors for electronic canopy characterization in vineyards: Advances to improve pesticide application methods, Sensors 11 (2) (2011) 2177–2194.

[57] F. Mazzetto, A. Calcante, A. Mena, A. Vercesi, Integration of optical and analogue sensors for monitoring canopy health and vigour in precision viticulture, Precision Agriculture 11 (6) (2010) 636–649.

[58] B. Grocholsky, S. Nuske, M. Aasted, S. Achar, T. Bates, A camera and laser system for automatic vine balance assessment, in: 2011 Louisville, Kentucky, August 7-10, 2011, American Society of Agricultural and Biological Engineers, 2011, p. 1.

[59] H. Moreno, C. Valero, J. M. Bengochea-Guevara, Á. Ribeiro, M. Garrido-Izard, D. Andújar, On-ground vineyard reconstruction using a LiDAR-based automated system, Sensors 20 (4) (2020). doi:10.3390/s20041102.

[60] M. Weiss, F. Baret, Using 3D point clouds derived from UAV RGB imagery to describe vineyard 3D macrostructure, Remote Sensing 9 (2) (2017) 111.

[61] A. I. de Castro, F. M. Jiménez-Brenes, J. Torres-Sánchez, J. M. Peña, I. Borra-Serrano, F. López-Granados, 3-D characterization of vineyards using a novel UAV imagery-based OBIA procedure for precision viticulture applications, Remote Sensing 10 (4) (2018) 584.

[62] A. J. Mathews, J. L. Jensen, Visualizing and quantifying vineyard canopy lai using an unmanned aerial vehicle (uav) collected high density structure from motion point cloud, Remote Sensing 5 (5) (2013) 2164–2183.

[63] V. González-Caballero, M.-T. Sánchez, J. Fernández-Novales, M.-I. López, D. Pérez-Marin, On-vine monitoring of grape ripening using near-infrared spectroscopy, Food Analytical Methods 5 (6) (2012) 1377–1385.

[64] I. Kalisperakis, C. Stentoumis, L. Grammatikopoulos, K. Karantzalos, Leaf area index estimation in vineyards from UAV hyperspectral data, 2D image mosaics and 3D canopy surface models, The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40 (1) (2015) 299.

[65] H. Aasen, A. Burkart, A. Bolten, G. Bareth, Generating 3D hyperspectral information with lightweight uav snapshot cameras for vegetation monitoring: From camera calibration to quality assurance, ISPRS Journal of Photogrammetry and Remote Sensing 108 (2015) 245–259.

[66] D. Dey, L. Mummert, R. Sukthankar, Classification of plant structures from uncalibrated image sequences, in: Applications of Computer Vision (WACV), 2012 IEEE Workshop on, 2012, pp. 329–336.

[67] R. Roscher, K. Herzog, A. Kunkel, A. Kicherer, R. Töpfer, W. Förstner, Automated image analysis framework for high-throughput determination of grapevine berry sizes using conditional random fields, Computers and Electronics in Agriculture 100 (2014) 148–158.

[68] R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, ORB-SLAM: a versatile and accurate monocular SLAM system, IEEE Transactions on Robotics 31 (5) (2015) 1147–1163.

[69] F. Schöler, V. Steinhage, Automated 3D reconstruction of grape cluster architecture from sensor data for efficient phenotyping, Computers and Electronics in Agriculture 114 (2015) 163–177.

[70] J. Mack, C. Lenz, J. Teutrine, V. Steinhage, High-precision 3D detection and reconstruction of grapes from laser range data for efficient phenotyping based on supervised learning, Computers and Electronics in Agriculture 135 (2017) 300–311.

[71] M. Klodt, K. Herzog, R. Töpfer, D. Cremers, Field phenotyping of grapevine growth using dense stereo reconstruction, BMC bioinformatics 16 (1) (2015) 143.

[72] F. Rist, K. Herzog, J. Mack, R. Richter, V. Steinhage, R. Töpfer, High-precision phenotyping of grape bunch architecture using fast 3D sensor and automation, Sensors 18 (3) (2018) 763.

[73] J. Mack, F. Schindler, F. Rist, K. Herzog, R. Töpfer, V. Steinhage, Semantic labeling and reconstruction of grape bunches from 3D range data using a new RGB-D feature descriptor, Computers and Electronics in Agriculture 155 (2018) 96–102. doi:https://doi.org/10.1016/j.compag.2018.10.011.

[74] V. Steinhage, F. Schöler, Automated reconstruction of 3D plant architecture applied to grapevine phenotyping, Informatik in der Land-, Forst-und Ernährungswirtschaft 2015 (2015).

[75] M. Gao, T.-F. Lu, Image processing and analysis for autonomous grapevine pruning, in: Mechatronics and Automation, Proceedings of the 2006 IEEE International Conference on, IEEE, 2006, pp. 922–927.

[76] H. Moreno, J. Bengochea-Guevara, A. Ribeiro, D. Andújar, 3D assessment of vine training systems derived from ground-based RGB-D imagery, Agriculture 12 (6) (2022). doi:10.3390/agriculture12060798.

[77] W. Yin, H. Wen, Z. Ning, J. Ye, Z. Dong, L. Luo, Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks, Frontiers in Robotics and AI 8 (2021). doi:10.3389/frobt.2021.626989.

[78] F. Marinello, A. Pezzuolo, F. Meggio, J. Martínez-Casasnovas, T. Yezekyan, L. Sartori, Application of the Kinect sensor for three dimensional characterization of vine canopy, Advances in Animal Biosciences 8 (2) (2017) 525–529.

[79] C. Neupane, A. Koirala, Z. Wang, K. B. Walsh, Evaluation of depth cameras for use in fruit localization and sizing: Finding a successor to kinect v2, Agronomy 11 (9) (2021). doi:10.3390/agronomy11091780.

[80] C. Hacking, N. Poona, N. Manzan, C. Poblete-Echeverría, Investigating 2-D and 3-D proximal remote sensing techniques for vineyard yield estimation, Sensors 19 (17) (2019). doi:10.3390/s19173652.

[81] A. Milella, R. Marani, A. Petitti, G. Reina, In-field high throughput grapevine phenotyping with a consumer-grade depth camera, Computers and Electronics in Agriculture 156 (2019) 293–306. doi:https://doi.org/10.1016/j.compag.2018.11.026.

[82] J. R. Rosell-Polo, F. Auat Cheein, E. Gregorio, D. Andújar, L. Puigdomènech, J. Masip, A. Escolà, Chapter three - advances in structured light sensors applications in precision agriculture and livestock farming, Vol. 133 of Advances in Agronomy, Academic Press, 2015, pp. 71–112. doi:https://doi.org/10.1016/bs.agron.2015.05.002.

[83] S. Nuske, K. Gupta, S. Narasimhan, S. Singh, Modeling and calibrating visual yield estimates in vineyards, in: Field and Service Robotics, Springer, 2014, pp. 343–356.

[84] S. Nuske, S. Achar, T. Bates, S. Narasimhan, S. Singh, Yield estimation in vineyards by visual grape detection, in: Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, IEEE, 2011, pp. 2352–2358.

[85] B. Xin, S. Liu, M. Whitty, Three-dimensional reconstruction of Vitis vinifera (L.) cvs Pinot Noir and Merlot grape bunch frameworks using a restricted reconstruction grammar based on the stochastic L-system, Australian Journal of Grape and Wine Research 26 (3) (2020) 207–219. doi:https://doi.org/10.1111/ajgw.12444.

[86] B. Xin, M. Whitty, A 3D grape bunch reconstruction pipeline based on constraint-based optimisation and restricted reconstruction grammar, Computers and Electronics in Agriculture 196 (2022) 106840. doi:https://doi.org/10.1016/j.compag.2022.106840.

[87] M. Herrero-Huerta, D. González-Aguilera, P. Rodriguez-Gonzalvez, D. Hernández-López, Vineyard yield estimation by automatic 3D bunch modelling in field conditions, Computers and Electronics in Agriculture 110 (2015) 17–26.

[88] G. Victorino, R. P. Braga, J. Santos-Victor, C. M. Lopes, Overcoming the challenge of bunch occlusion by leaves for vineyard yield estimation using image analysis, OENO One 56 (1) (2022) 117–131. doi:10.20870/oeno-one.2022.56.1.4863.

[89] K. W. Eccleston, I. G. Platt, A. E.-C. Tan, SAR for grape bunch detection in vineyards, in: Microwave Symposium (AMS), 2018 Australian, IEEE, 2018, pp. 3–4.

[90] D. Henry, H. Aubert, P. Galaup, T. Véronèse, Dynamic estimation of the yield in precision viticulture from mobile millimeter-wave radar systems, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15. doi:10.1109/TGRS.2021.3133017.

[91] M. Legg, S. Bradley, Ultrasonic arrays for remote sensing of pasture biomass, Remote Sensing 12 (1) (2020). doi:10.3390/rs12010111.
URL https://www.mdpi.com/2072-4292/12/1/111

[92] S. Chen, Y. Li, N. M. Kwok, Active vision in robotic systems: A survey of recent developments, International Journal of Robotics Research 30 (11) (2011) 1343–1377.

[93] B. Parr, M. Legg, F. Alam, Analysis of Depth Cameras for Proximal Sensing of Grapes, Sensors 22 (11), dio:10.3390/s22114179 (2022).
URL https://www.mdpi.com/1424-8220/22/11/4179

[94] B. Parr, M. Legg, F. Alam, Grape yield estimation with a smartphone's colour and depth cameras using machine learning and computer vision techniques, Computers and Electronics in Agriculture 213 (2023) 108174. doi:https://doi.org/10.1016/j.compag.2023.108174.
URL https://www.sciencedirect.com/science/article/pii/S0168169923005628

[95] B. Parr, M. Legg, S. Bradley, F. Alam, Occluded grape cluster detection and vine canopy visualisation using an ultrasonic phased array, Sensors 21 (6) (2021). doi:10.3390/s21062182.
URL https://www.mdpi.com/1424-8220/21/6/2182

[96] B. Parr, M. Legg, F. Alam, S. Bradley, Acoustic identification of grape clusters occluded by foliage, in: Sensors and Applications Symposium (SAS 2020), Kuala Lumpur, Malaysia, 2020, pp. 1–6. doi:10.1109/SAS48726.2020.9220078.

[97] M. Tölgyessy, M. Dekan, Ľ. Chovanec, P. Hubinský, Evaluation of the Azure Kinect and its comparison to Kinect V1 and Kinect V2, Sensors 21 (2) (2021). doi:10.3390/s21020413.

[98] F. Schöler, V. Steinhage, Towards an automated 3D reconstruction of plant architecture, in: International Symposium on Applications of Graph Transformations with Industrial Relevance, Springer, 2011, pp. 51–64.

Chapter 2

Analysis of 3D Cameras for Reconstructing Spherical Clusters

This chapter presents a draft of a paper that has not yet been published. It has been included on the basis that we believe it to be novel work and its inclusion provides context to our understanding of 3D camera performance explored in later chapters.

Analysis of 3D Cameras for Reconstructing Spherical Clusters

Baden Parr, Mathew Legg*, Fakhrul Alam

Department of Mechanical and Electrical Engineering, Massey University, Auckland, New Zealand

Abstract

Commercial, low-cost, 3D cameras that generate real-time 3D scans have potential for use in grape yield estimation. However, their suitability for this task has not been thoroughly explored. Correctly capturing the surface curvature concave details between neighbouring grapes details will be important for accurate yield estimation. Traditionally, 3D cameras have been evaluated using flat targets, such as walls or boxes, and there is a lack of research evaluating their ability to accurately capture the detail of groups of spherical objects. This paper investigates the performance of three different types of depth cameras to correctly capture reference spherical objects and presents the accuracy of each with respect to surface normals and camera perspective. Furthermore, their ability to reconstruct concave details present between neighbouring spheres and the impact of distance on accuracy is evaluated. The results indicate that the Microsoft Kinect V2 time of flight based camera is the most suitable, with average projected errors of less than 2mm across all incidence angles up to 1 meter.

Keywords: grape yield estimation, 3D camera, 3D camera performance, Kinect, Intel D415

1. Introduction

In recent years, low-cost real-time 3D cameras have been introduced to the market. Several recent studies have indicated that these may be suitable for use in vineyards. Marinello et al. demonstrated the use of a first generation Microsoft Kinect for the volumetric assessment of vine canopies [1, 2] and Gao also used it to analyse vine structure for the purpose of automated pruning [3]. However, some researchers have speculated that 3D cameras are not currently able to provide enough precision for accurate grape cluster analysis [4].

There are many reported studies that have evaluated the accuracy of 3D cameras [5, 6, 7, 8, 9, 10, 11]. However, these generally use solid flat objects as targets, e.g. walls or boxes. Additionally, there is often a focus on distance accuracy [12] and characterisation of the various sources of noise that affect 3D cameras [13, 14, 15]. The working group, VDI/VDE-GMA, proposed guidelines for assessing the performance of optical

*Corresponding author

Email addresses: 1badenparr@gmail.com (Baden Parr), M.Legg@massey.ac.nz (Mathew Legg), F.Alam@massey.ac.nz (Fakhrul Alam)

Table 1: A summary of the 3D camera technologies that have been analysed in this study.

Technology	Approach	Comment
Structured Light	Distortion of a known pattern	Requires projecting a pattern of light into a scene, more robust than other methods but is impacted by natural light.
Passive Stereo	Disparity between image pairs	Disparity can be difficult to find in regions with low contrast. Works well in direct sunlight but adding more than 2 cameras increases processing exponentially.
Active Stereo	Disparity between image pairs assisted by projected light	Utilises a combination of structured light and traditional stereo to assist in regions with low contrast.
Time of Flight	Time taken for light to return	Requires precise timing, accuracy can be impacted by temperature as well as the spectral properties of the object of interest.

Table 2: A summary of the 3D cameras that have been analysed in this study.

Camera	Technology	Field of View (h x v)	Pixel Resolution
Microsoft Kinect V1	Structured Light	58.5 x 46.6	640 x 480 pixel
Microsoft Kinect V2	Time-of-Flight	70.6 x 60	512 x 424 pixels
Intel RealSense D415	Active/Passive Stereo	69.4 x 42.5	1280 x 720 pixels

3D measuring systems [16]; a procedure that has since been employed for the analysis of popular commercial 3D cameras [17]. This standard focuses on absolute accuracy and "error of length" measurements. There is a need for further work to investigate if commercial 3D cameras are suitable for grape yield estimation. Grape yield estimation poses unique challenges for 3D cameras due to the grape's size and cluster complexity. For this application, a key requirement is that individual grapes are able to be identified, slight errors in the exact distance of the grape from the camera is of less importance. To detect a small spherical object such as a grape, a 3D camera must be able to resolve the slight differences in distance between neighbouring points that describe the objects curvature. This requires that the resolution and accuracy of the scan must be enough so that individual grapes are identifiable. Additionally, grapes are approximately spherical and grow in dense clusters. It is important to be able to identify the convex regions between neighbouring grapes.

This paper investigates the performance of three different 3D cameras: the first generation Microsoft Kinect structured light 3D camera (V1), the second generation Microsoft Kinect Time of Flight (ToF) 3D camera (V2), and the Intel RealSense D415 active/passive stereo 3D camera. These are operated using four common 3D camera technologies: structured light, ToF, stereo, and a hybrid of structured light and stereo, refer to Table 1. Table 2 provides key properties of the three 3D cameras used in this study.

The focus of this work is determining the accuracy of these cameras for the application of grape yield estimation. This is achieved by using small spherical proxies as reference targets, both singularly and arranged small clusters. These are used to benchmark each camera's performance across several parameters

deemed vital for accurate identification of grape berries. The results from this study will also be relevant to many other applications where scans of fine detail 3D objects are required. In Section 2, the camera's expected performance for imaging grapes due to properties such as field of view and resolution are discussed. Section 3 then describes the methodology used to capture and process the data. Experimental results showing the performance of the cameras for a single sphere and multiple spheres are then presented respectively in Sections 4 and 5. Finally a discussion and conclusion are provided in Section 6.

2. Pertinent 3D Camera Parameters

The properties of 3D cameras such as resolution, field of view, and operational distance have significant impact on the performance of 3D cameras. This section takes a theoretical look at how these parameters effect the number of depth pixels that can be used to image an object and the potential sources of error in the resulting depth map.

2.1. Pixel density

The ability of a 3D camera to capture the shape of a grape accurately will be related to the number of depth pixels that fall across its surface. This will depend on the 3D camera's resolution, field of view, and distance from the object's surface, (please see Figure 1). The number of pixels P_n that may be used to image an object with a visible area A can be expressed as

$$P_n = P_d(z)\ A \tag{1}$$

where $P_d(z)$ is the pixel density (number of depth pixels per square mm) at a distance z from an object. If we approximate the visible area of a grape as a circle with a radius r, then the number of pixels that will resolve the grape's surface can be computed as

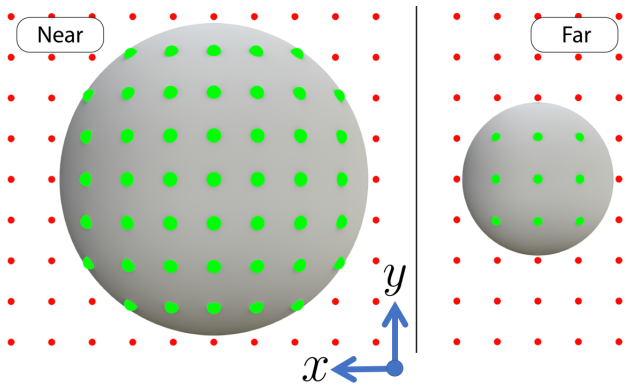


Figure 1: Diagram illustrating how the number of depth pixels used to image an object changes with imaging range. This can be express as a distance dependant spatial pixel density.

$$P_n = P_d(z)\ \pi r^2. \tag{2}$$

The pixel density can be calculated using

$$P_d(z) = \frac{R_x}{\gamma_x(z)}\ \frac{R_y}{\gamma_x(z)} \tag{3}$$

where R_x and R_y are the 3D camera's horizontal and vertical pixel resolutions and $\gamma_x(z)$ and $\gamma_y(z)$ are respectively the spatial horizontal and vertical field of view (in mm) at given distance. These fields of view may be calculated using

$$\begin{aligned} \gamma_x(z) &= 2\tan\left(\frac{h}{2}\right)\ z \\ \gamma_y(z) &= 2\tan\left(\frac{v}{2}\right)\ z \end{aligned} \tag{4}$$

where h and v are respectively the camera's angular horizontal and vertical field of view in degrees. Values for each camera's angular field of view and resolution are given in Table 2.

The pixel densities of the cameras used in this study have been computed for distances up to 1.5 m and are shown in Table 3. These can be used to estimate the number of pixels that could be used to image a grape of a particular size at a given distance. For example, a 20 mm diameter grape positioned a meter from the Intel D415 3D camera would resolve it with roughly 270 depth pixels. In comparison, the Kinect V1 and V2 at the same distance will see it with roughly 100 depth pixels and 40 depth pixels respectively. If the grapes are too far from the sensor, there is a strong likelihood that there will not be enough pixels to successfully describe its surface. In addition to the number of depth pixels that can be used to image an object, the accuracy of these depth pixels is also important. Noise characteristics of the different 3D cameras are factors that needs to be considered.

2.2. Depth Quantisation Error

Depth quantisation is one of the potential sources of error for 3D cameras. The depth values generated by a 3D camera will be quantised into discrete depth distances. This is illustrated in Figure 2. This depth quantisation is roughly linear with distance for many cameras such as the Microsoft Kinect V2, which uses time of flight technique. However, for 3D cameras that rely on triangulation to determine per pixel depth values, the depth resolution is proportional to the pixel density at a certain depth. This results in the non-linear quantization of depth values within a scene. For the structured light based Microsoft Kinect V1, this non-linearity has a substantial impact within the expected working range of a grape yield estimation system. The relationship between the pixel disparity Δp and the measured depth distance $\bar{z}$ is given as

Table 3: The point density of the analysed cameras against increasing depths.

	Pixel Density ($pixels/mm^2$)		
Distance (mm)	**Microsoft Kinect V1**	**Microsoft Kinect V2**	**Intel RealSense D415**
500	1.27	0.53	3.42
600	0.88	0.37	2.38
700	0.65	0.27	1.75
800	0.50	0.21	1.34
900	0.39	0.16	1.06
1000	0.32	0.13	0.86
1100	0.26	0.11	0.71
1200	0.22	0.09	0.59
1300	0.19	0.08	0.51
1400	0.16	0.07	0.44
1500	0.14	0.06	0.38

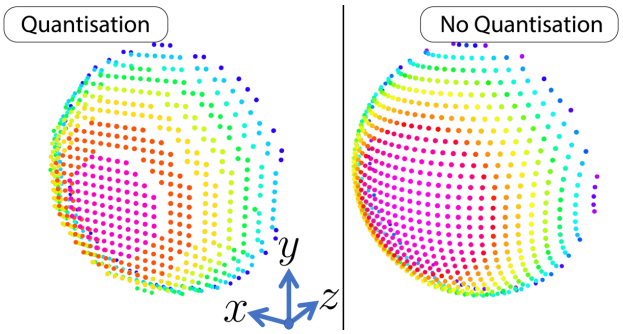


Figure 2: Graphical diagram demonstrating the effect of depth quantisation on a spherical surface.

$$\bar{z} = \frac{f\ b}{\Delta p} \tag{5}$$

where f is the focal length (in pixels) and b is the optical base line (in mm) between the light projector and the IR camera. For the Kinect V1, these values are 580 pixels and 75 mm respectively. To increase depth resolution, cameras will typically calculate a sub-pixel disparity and utilise a disparity offset to shift the available bit depth towards more usable depths ranges.

The Kinect V1 is able to identify disparity with a sub-pixel rate of 8. To correct for this, the disparity must also be scaled by 1/8 to calculate the distance. The complete relationship for the Kinect V1 is therefore given by

$$\bar{z} = \frac{f\ b}{1/8\,(p_{offset} - \Delta p)} \tag{6}$$

where p_{offset} is the offset value unique to every camera but will be around 1090 pixels [18]. With this

information, the quantisation of depth can be established as the change in depth distance $\Delta\bar{z}$ per integer change in pixel disparity Δp. i.e. what is the change in depth distance between Δp and $\Delta p + 1$. This can be calculated by taking the derivative of (6) with respect to Δp giving

$$\Delta\bar{z} = \frac{8f\ b}{p_{off}^2 + \Delta p^2 - 2\ p_{off}\ \Delta p + p_{off} - \Delta p}. \tag{7}$$

The result is a polynomial relationship between the disparity Δp and its equivalent depth quantisation $\Delta\bar{z}$. For a typical Kinect V1, this will result in a depth quantisation of roughly 0.7 mm at a range of 0.5 m increasing to around 6.5 mm at 1.5 m. For a grape that is approximately spherical with a diameter of 20 mm, this equates to roughly 14 depth values to describe the visible surface when 0.5 m from the camera. Furthermore, at a distance of 1.5 m, the same grape will only be resolved by a single depth value making it impossible to determine surface curvature.

Although the Intel D415 uses stereo disparity, it has reduced the impact of depth quantisation errors though improvements in sub pixel disparity matching algorithms, increased sensor resolution, and dynamic offset adjustment [19]. The Kinect V2 uses time of flight and does not have this same quantisation error issue. Our measurements with the Kinect V2 indicated that its quantisation error was approximately constant at 1 mm for all distances measured.

2.3. Other Noise

There are a number of other types of noise that can be present on 3D camera data. These include noise related to flying pixels, surface incident angle, and the transparency of the surface being imaged. However, these are more difficult to quantify numerically. Experimental measurements are therefore needed to investigate how these will affect the performance of these 3D cameras for imaging grapes.

3. Experimental Methodology

3.1. CNC Measurement Platform

A two axis CNC gantry was utilised to accurately and reliably position the 3D cameras as shown in Figure 3. Custom camera mounts were designed to ensure the optical centres of each camera were positioned in the same relative position when mounted to the CNC. The gantry features an absolute accuracy of 0.01 mm and a moveable area of 1.5 m × 1.5 m. Measurements were conducted sequentially and with the same lighting across all cameras.

3.2. Reference Sphere

To approximate the surface curvature of a grape, a spherical reference object was chosen to evaluate the performance of each 3D camera. A painted table tennis ball (40 mm diameter) was used for this purpose,

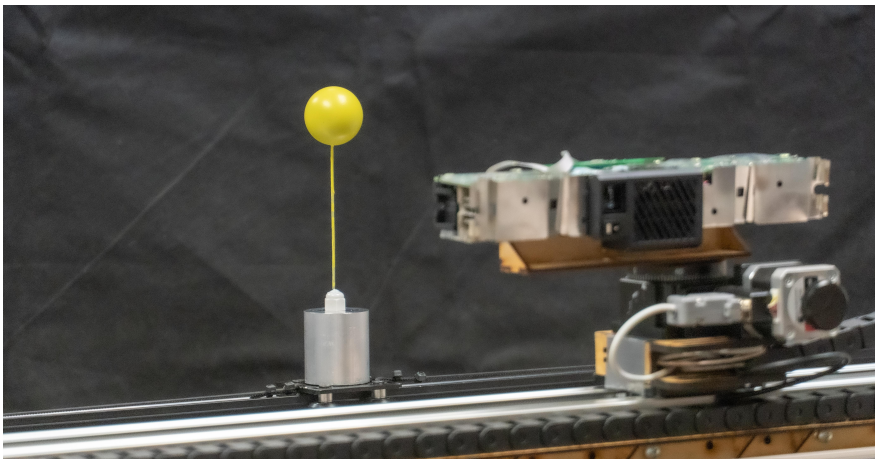


Figure 3: Experimental set-up showing a 3D camera (Kinect V2) mounted to a computer controlled measurement platform positioned in front of the reference sphere.

(please see Figure 3). Although this is much larger than a grape, it was decided that the larger size would reduce the potential effect of depth quantisation noise and hence better allow the accuracy of the cameras for imaging curved surfaces to be evaluated. To be a more accurate proxy of a real grape, it was desirable for the reference sphere to exhibit similar surface characteristics. To achieve this, both the colour and the reflectivity were taken into account. The colour of ripe green table grapes was matched with assistance from a professional painter. The reflectivity of the reference was also matched as several of the 3D camera technologies have the potential to be affected by spectral highlights. This was achieved by comparing reflectance profiles of various sheens of paint with that of grapes in controlled lighting conditions. However, it should be noted that the grapes are likely to also have some translucent properties while the reference sphere is opaque.

3.3. Measurement Procedure

Each camera was aligned so that the reference sphere was centrally located within the frame to reduce adverse effects resulting from any optical distortion from the lens. Each scan was captured against a black background to make segmenting the point cloud easier. All cameras were configured to default settings to assess their performance without adjusting parameters to optimise results for our given environment. Point clouds from the Kinect V1, Kinect V2, and Intel D415 were captured using KScan3D v1.2, MATLAB R2018a, and the Intel RealSense Viewer v2.24.0 respectively. The Intel D415 performed the role of two technologies, active and passive stereo. This was achieved by disabling the projector for the passive stereo tests.

Each 3D camera technology was used to capture point clouds at 0.1 m incremental distances between 0.5 m and 1.5 m away from the centre of the reference sphere. At each distance, 20 frames were taken to

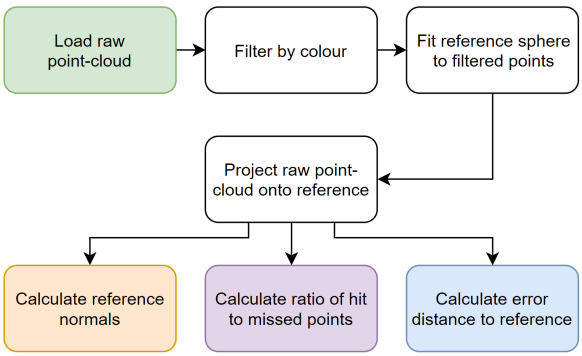


Figure 4: Overview of the process used to analyse each 3D camera's performance against the single sphere reference.

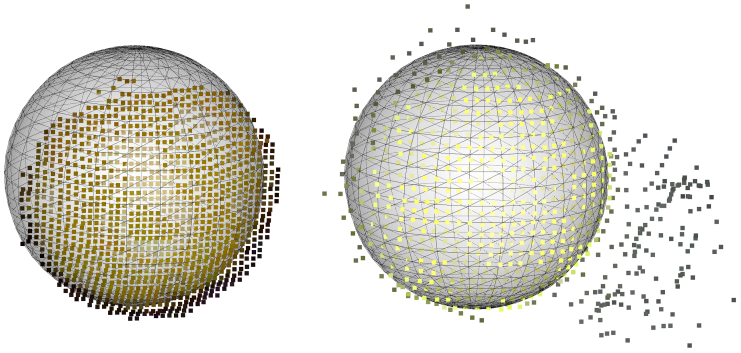


Figure 5: Raw point clouds from the Kinect V1 (left) and Kinect V2 (right) after alignment with the virtual reference sphere.

explore the statistical variation at that range. This distance range was chosen as it reflects the expected operating conditions if these cameras were to be used in a vineyard environment and is within the operating range of each camera. Assuming the cameras are mounted to farm equipment that traverses the rows, it would be difficult to position them outside of this range.

3.4. Post-processing of Point Clouds

The point cloud for each frame was processed using the method outlined in Figure 4. Point clouds were segmented and filtered by colour to isolate the points belonging to the reference sphere. It was found that a simple green channel threshold worked the best for this as it had a very distinct bimodal distribution. However, as the RGB sensor in each 3D camera had different colour accuracies, this threshold needed to be adjusted for each device. A threshold value of 40 was used for the Kinect V1, 150 for the Kinect V2, and 100 for the Intel D415. Points over this threshold were treated with high confidence as belonging to the reference sphere. After the scans had undergone colour thresholding, the remaining points are aligned to a digital reference sphere (or group of spheres) using an Iterative Closest Point (ICP) algorithm [20], (please see Figure 5).

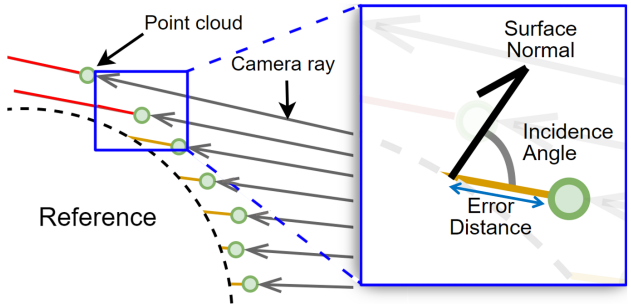


Figure 6: Graphical diagram showing the principle used to calculate the error and incidence angle with respect to a reference surface. In addition, red lines represent projections that have missed the reference, while orange lines represent hits.

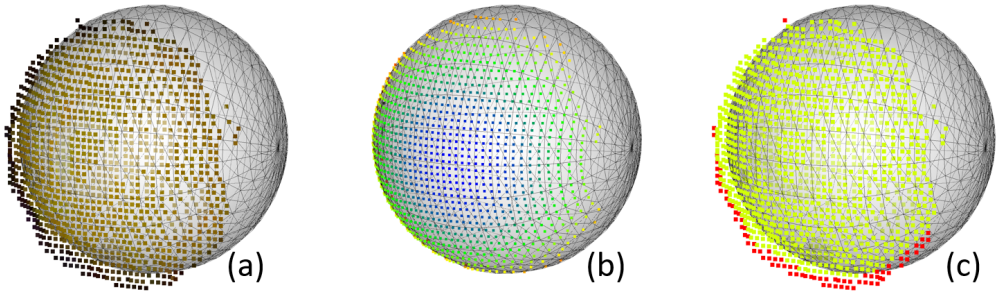


Figure 7: Plots illustrating the processing steps used to calculate the distance error and flying pixels for a 3D capture from the Kinect V1 at 0.6 m. Plot (a) shows the aligned point cloud, (b) the points where the rays projected from the camera through the depth points intersect the reference sphere, and (c) the classified points that hit (yellow) or missed the reference (red).

3.4.1. Calculation of Distance Error

Once the point clouds have been aligned with the virtual reference sphere, the error for each point can be calculated between the captured point cloud and reference surface. This could have been achieved by calculating the distance from the measured point to the closest point on the reference surface as is often done in literature [5]. However, this naively produces lower errors in areas where the camera ray hits the surface at a high incidence angle. Instead, distance errors were calculated using a ray that was projected from the camera location through the measured point. If the ray hit the reference sphere, the distance error was then obtained by measuring the distance along the ray from the measured point to the surface of the reference sphere, (please see Figures 6 and 7).

3.4.2. Calculation of Incident Angle

The incidence angle was obtained for each measured point that had a ray that intersected with the reference sphere. At each of these locations, the surface normal was calculated by normalising the vector from the centre of the reference sphere to the point of intersection. The incidence angle θ_i for each point is then calculated using the dot product of the two vectors as

$$\theta_i = \cos^{-1}\left(\hat{R}_i \cdot \hat{S}_i\right) \tag{8}$$

where $\hat{R}_i$ is the normalised camera ray vector for the i^{th} point, and $\hat{S}_i$ is the normalised surface normal at the point where $\hat{R}_i$ intersects the reference surface.

4. Single Sphere Results

4.1. Average Error vs Distance

Absolute errors were calculated for every point using the projection method discussed previously. Example point clouds with distance errors for each camera are shown in Figure 8. Each of the 20 point clouds were processed independently and the results used to estimate temporal statistics for a given measurement distance. This is a useful indication of the repeatability of the measurements of each camera. The average distance errors are presented in the box plot shown in Figure 9.

The Kinect V2 time of flight camera shows relatively uniform average errors over all tested distances with average errors constantly less than 2 mm up a distance of 1 m. Past 1 m, its performance and repeatability begins to decrease. Interestingly, both the structured light Kinect V1 and the active stereo Intel D415, performed almost identically. The average errors for both steadily increases with measurement distance from close to 2 mm up to around 5 mm. Passive stereo featured the worst performance in comparison to the other technologies. At close range, the average error was close to 3 mm, increasing slightly to an average of 6 mm at the furthest distance. It is also worth noting the significant levels of variance between consecutive scans observed at these distances.

4.2. Erroneous Points

The number of flying pixels can be estimated by using the projection method described above. Points close to the reference surface that that had projected rays from the camera through the depth pixel missing the reference sphere were assumed to be "flying pixels" surrounding the object of interest, (please see Figures 6 and 7). Taking the ratio of hit to miss points normalises for the unique point density of each camera so each camera can be directly compared. The results can be seen as bar charts in Figure 10. The charts show the ratio distribution of 20 frames captured at each distance. Significant insight can be drawn from these results. Firstly, the Kinect V2 at close range has a considerable proportion of missed points. At a distance of 0.5 m, the number of missed points is greater than the number of hit points indicating significant levels of flying pixels surrounding the reference sphere.

In addition to the hit and miss ratio, the distribution of missed points is also a key indicator of performance. Missed points that are close to the object of interest are likely going to have a different impact on subsequent point cloud processing compared to those that more wide spread. To analyse this, the distance

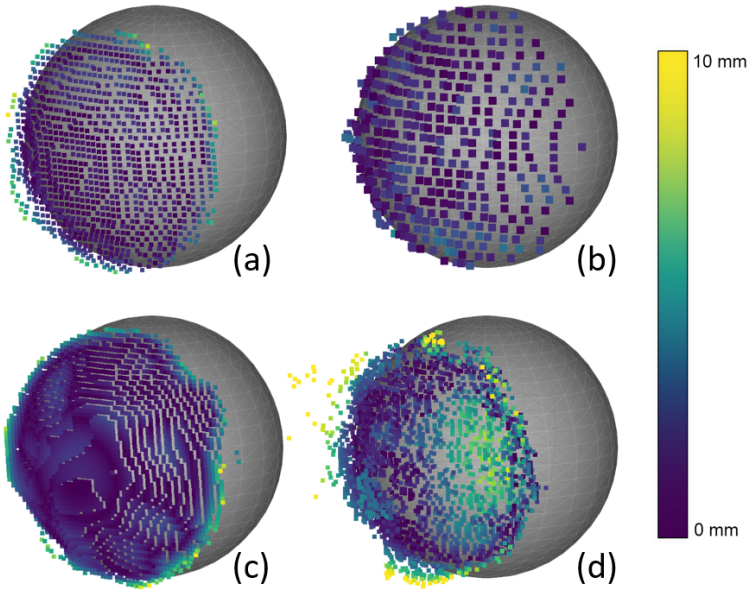


Figure 8: Example point clouds from each of the four 3D camera technologies captured at a distance of 0.6 m: (a) the Kinect V1, (b) Kinect V2, (c) Intel D415 active stereo, and (d) Intel D415 passive stereo. Points have been coloured according to their projected error metric.

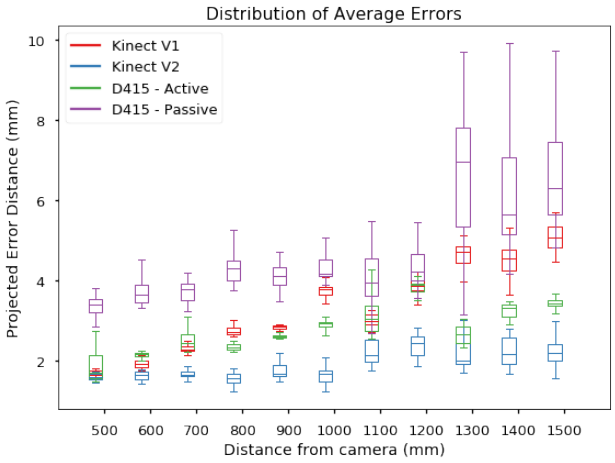


Figure 9: Box and whisker plots of the average errors at each measurement distance for the four tested 3D camera technologies.

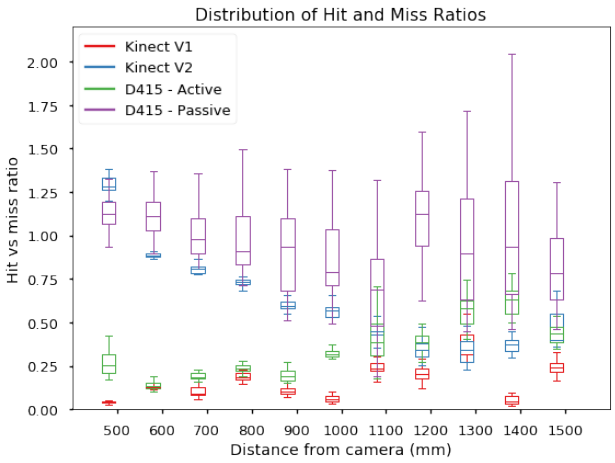


Figure 10: Box and whisker plots of the hit verses miss ratios at each measurement distance for the four tested 3D camera technologies.

from each missed point to the surface of the reference sphere can be calculated. For each camera, the distribution was evaluated at three primary distances, 0.5 m, 1.0 m, and 1.5 m. The results are presented as Cumulative Distribution Functions (CDF) in Figure 11.

The CDF for Kinect V2 shows that the distribution of missed points gets closer to the surface as the distance between the camera and the sphere increases. This is potentially an artefact of the cameras geometry since less of the reference sphere is seen as illuminated by the ToF projector the closer the object is to the sensor. However, the opposite is observed for the Kinect V1 where the performance gets worse with increased distance. The Intel D415 in active mode also follows this trend, albeit relatively loosely. The poor performance for D415 in passive mode can be clearly observed for all three distances.

4.3. Influence of Incidence Angle

The orientation of the surface relative to the camera for each pixel can be used to analyse the relationship between surface angle and potential error distance. This is an important performance factor for the detection of small spherical objects such as grapes. The more grape surface curvature that can be accurately detected, the more information will be available to be used in subsequent algorithms. This is especially vital for larger measurement distances or low resolution cameras where the point density is relatively low.

The incidence angle was calculated for all points that hit the surface of the virtual reference when projected. This was repeated for all 20 captures taken at each distance. Figure 12 shows the resulting scatter plot of the incidence angle vs error distance for the Kinect V1 and Kinect V2 depth cameras at a measurement distance of 0.6 m. Clear distinctions can be made from these results. The Kinect V2 has a relativity uniform error distribution across all incidence angles with the majority of errors falling below 2 mm. In contrast, the Kinect V1 has the smallest errors near zero degrees and is relativity uniform until 60

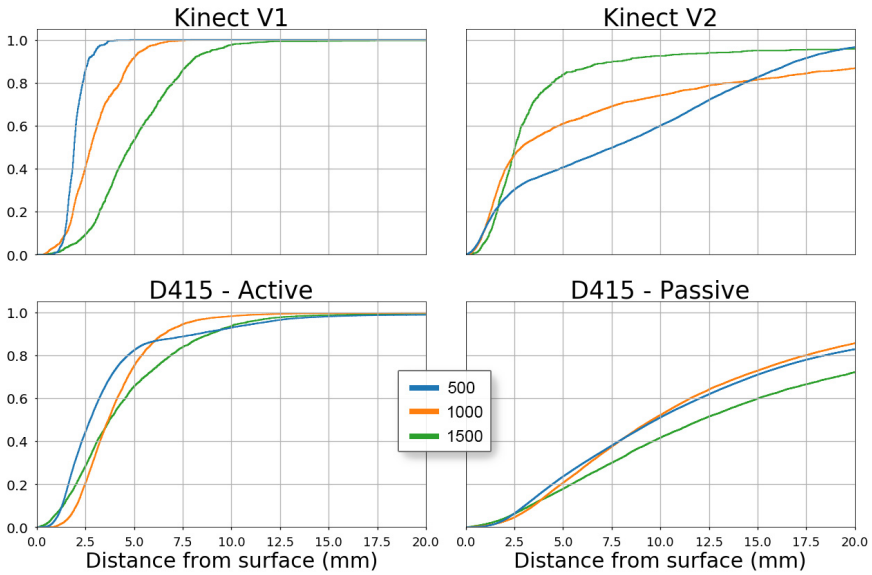


Figure 11: CDF for the distance of missed points from the surface of the reference sphere for each tested 3D technology at three measurement distances. The X-axis has been made uniform for ease of comparison.

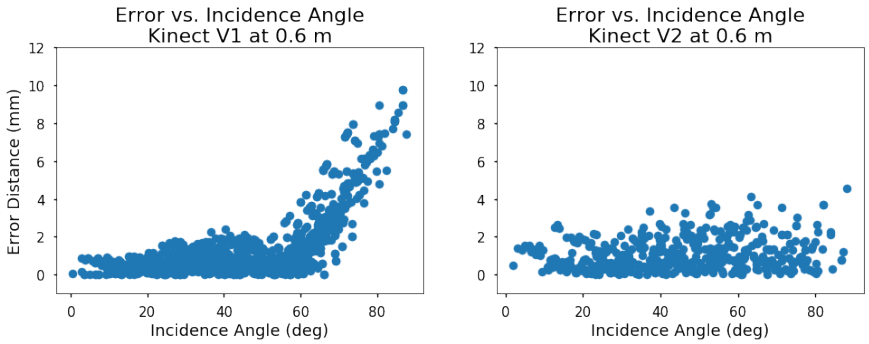


Figure 12: Distance error as a function of incidence angle.

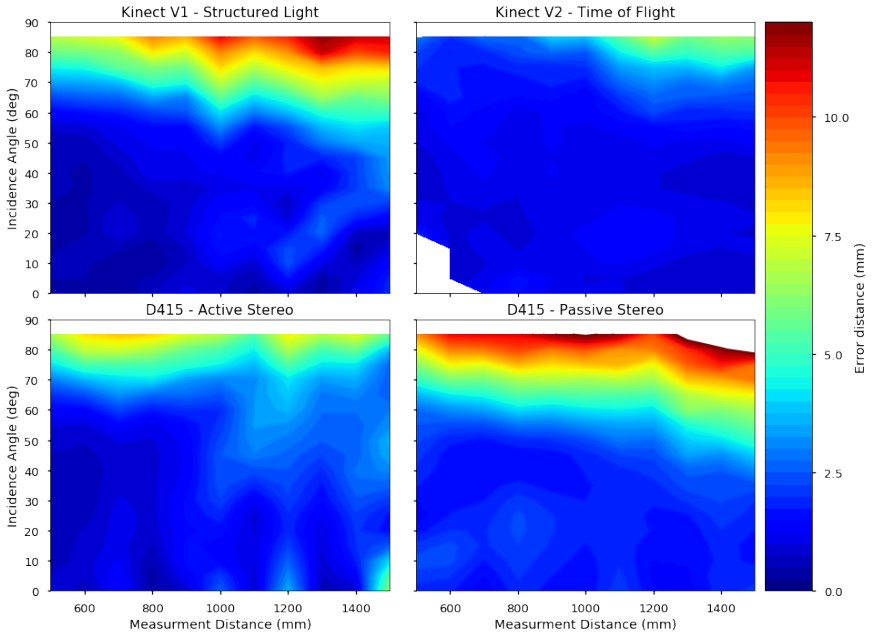


Figure 13: Heat maps showing the distribution of mean errors with respect to camera-surface incidence angles for the tested cameras of a single 40 mm sphere.

degrees where it rapidly degrades.

To visualise the data across all measurement distances, the errors from each scan were collated into discrete bins five degrees wide. The mean for all errors in each bin were computed and the result displayed as a heat map of the incidence angle verses measurement distance where the mean error is mapped to the colour. The results are shown in Figure 13. This visualisation again shows the uniformity of the Kinect V2's angular performance up to 1 m, beyond which it starts to degrade slightly. Both the Kinect V1 and Intel D415 cameras show poorer performance, particularly the D415 when operated in passive stereo mode.

5. Surface Analysis of Groups of Spheres

Individual grapes can be adequately approximated by spheres. However, grape clusters are complex structures that feature significant concave regions between neighbouring grapes. These can be challenging for 3D cameras to adequately capture. The sharpness of these features could be lost due to internal 3D camera filtering algorithms that try to reduce noise. To assess the ability of 3D cameras to capture these regions, measurements were conducted using a tight grouping of three spheres.

A group of three spheres were created using the surface finish discussed in Section 3.2. Point clouds were captured by each camera at distances between 0.5 m and 1.5 m in steps of 0.1 m. At each distance, 20 point clouds were captured from each camera for further statistical evaluation. After capture, each point cloud

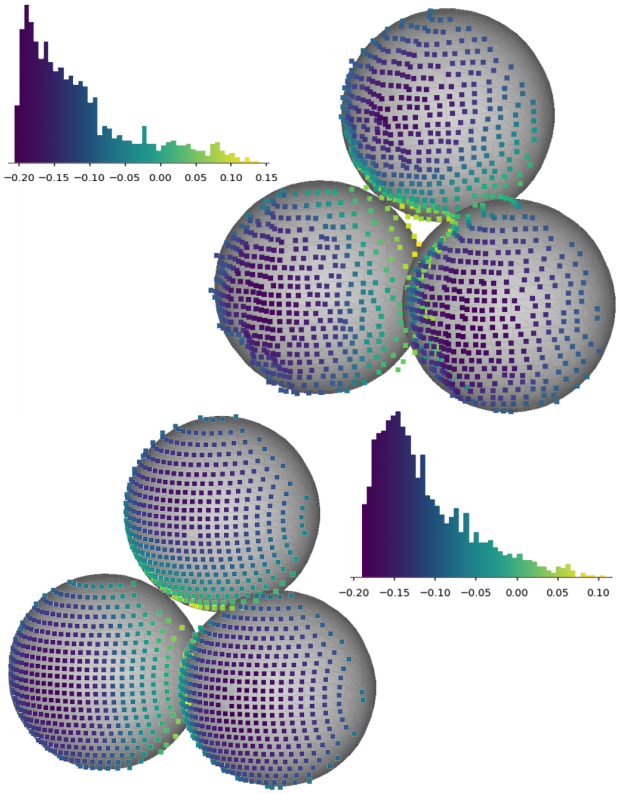


Figure 14: Processed point clouds coloured according to each points convergence metric. The top point cloud shows the raw capture from the Kinect V2 at a distance of 0.6 m. The bottom plot shows the projected locations of these points on the reference spheres, which corresponds to the ideal (no noise) case.

underwent colour thresholding and subsequent ICP alignment to a virtual reference of the sphere group. Errors for each point cloud are computed as before using the camera projection method shown in Figure 6.

A concave region in the point cloud can be expected where two or more spheres touch. To identify these regions, a convergence metric was calculated for each point as the average angle between the point's normal and the vectors to its k nearest neighbouring points. For a group of k neighbours around a selected point $\hat{X_n}$, the convergence value is calculated as

$$\gamma[n] = \frac{\sum_{i=1}^{k} \hat{\eta_n} \cdot \frac{\hat{X_n} - \hat{S_i}}{|\hat{X_n} - \hat{S_i}|}}{k} \tag{9}$$

where $\hat{\eta_n}$ is the normal vector of the n^{th} point, and $\hat{S_i}$ is the coordinate of the i^{th} neighbour. This process results in a convergence value between negative and positive one for each point in the point cloud. Negative values correspond to points in convex regions and positive values correspond to concave regions. As most of the visible regions are convex, the resulting distributions are heavily weighted to values less than zero, see Figure 14.

After projection, a convergence analysis was performed on the raw point values in addition to the

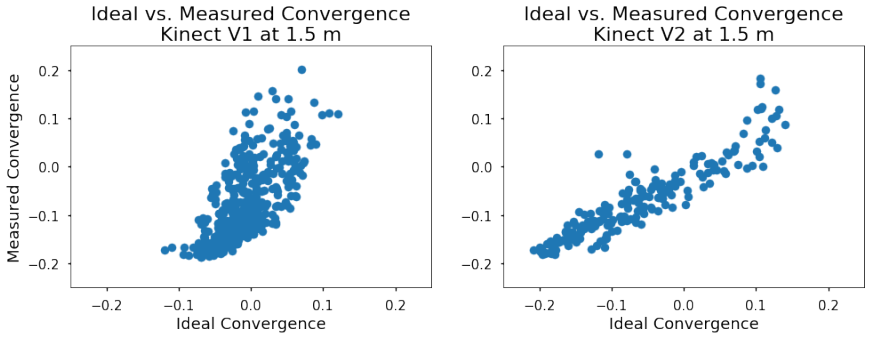


Figure 15: Scatter plots of the ideal and measured convergence values for point clouds captured by the Kinect V1 (left) and Kinect V2 (right) at a distance of 1.5 m. R-values of 0.63 and 0.90 were achieved respectively.

projected point locations on the reference spheres. These projected locations represent where the point would be if the 3D camera had no error. This is done to establish what the true convergence value should be for each point vs what is actually presented by the camera. In this way, a normalised comparison can be achieved between each camera that is independent of the pixel density.

5.1. Results

An example of the convergence analysis performed for the Kinect V2 is shown in Figure 14. The top point cloud shows the raw points coloured based on their convergence metric. The second shows the same process applied to ideal projected points lying on the surface of the reference spheres. Positive values, shown in yellow, are mapped to concave regions between the spheres, while dark blue negative values make up the majority of rest of the surface. It can be seen that the convergence distributions between both raw and projected point clouds are very similar. This is due to the ability of the Kinect V2 to capture surface features accurately.

The relationship between each of these distributions can be further explored by means of simple linear regression. Figure 15 shows scatter plots for point clouds taken by the Kinect V1 and the Kinect V2 at a distance of 1.5 m. As can be seen from the graphs, the Kinect V2 shows a strong relationship between the ideal convergence of the projected points vs that of the raw point cloud achieving a correlation coefficient R-value of 0.9. In comparison, the Kinect V1 struggles to remain consistent with the ideal values as confirmed by its low R-value of 0.63.

The correlation coefficient was computed in this way for each point cloud captured. At each distance, the average R-value was calculated from the 20 corresponding frames. The results can be seen in Figure 16. The Kinect V2 gave the best performance retaining R-values above 0.9 for the entire measurement range. The Kinect V1 and D415 used in active mode performed comparably but the performance dropped in comparison as the range increased. The D415 in passive stereo mode again performed very poorly for all distances.

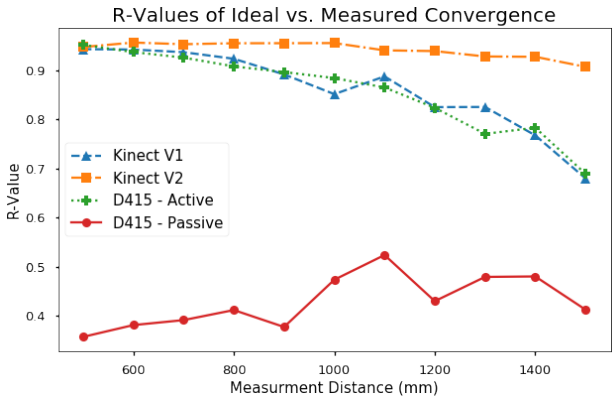


Figure 16: R-Values computed for all four cameras at each tested depth between 0.5 m and 1.5 m.

6. Discussion and Conclusion

For grape yield estimation, 3D cameras must be able to describe the spherical surface of grapes accurately enough to identify the size and position of individual berries within a cluster. To achieve this, it is crucial for 3D cameras to precisely describe the entire visible surface of a grape at all incidence angles. In addition, clear distinction between neighbouring grapes within a point cloud is necessary. Accurate representation of concave regions between grapes is vital for this purpose. This study has looked at four traditional 3D camera technologies as provided by three consumer 3D cameras frequently used within literature. Of the cameras and technologies tested, the Microsoft Kinect V2 ToF camera was found to be significantly better than the alternative technologies in many of the tests. Up to distances of 1 m, the Kinect V2 was able to describe the surface of the reference across all incidence angles with an average projected error of less than 2 mm. In contrast, the structured light and stereo cameras struggled to accurately assess incidence angles past 60 degrees. Additionally, the Kinect V2 demonstrated an excellent ability to capture concave regions between groups of spheres. Compared to the ideal convergence assessment, the Kinect V2 was able to achieve R values in excess of 0.9 across all measurement distances. With regards to spurious flying pixels, the Kinect V2 demonstrated poor performance at close range but improved significantly as distance increased. However, due to the nature of how the ToF camera captures these points, it is expected that they can be filtered in a post processing stage by looking at the per point variance across consecutive frames.

One of the biggest limitations of the Kinect V2 appears to be its pixel density. For an approximately spherical grape, 10 mm in diameter, it will be resolved by roughly 10 pixels at a distance of 1 m. This can be compared to the Kinect V1 and Intel D415 which will resolve the same object with 25 pixels and 67 pixels respectively. However, despite this, it is worth considering that the ability to trust a limited number of pixels is more valuable than having more pixels that cannot be trusted at all. Continuing with the same example, the Kinect V2, Kinect V1, and D415 in active mode achieved average distance errors of 1.6 mm,

3.7 mm, and 4.4 mm respectively. For a 10 mm diameter grape, distance errors that are nearly half its radius will make determining its true surface curvature very difficult. The results from this study have clearly highlighted that the time of flight 3D camera technology shows the most promise as a tool for grape yield estimation.

This study made extensive use of a 40 mm reference sphere which is considerably larger than a typical vine grape. Realistically, mature grapes range in size from 10 to 20 mm in diameter. It is expected that the results presented in this paper for the larger reference will be directly comparable to the realistically sized grapes by utilising known pixel densities and field of views of the tested cameras, (please see Table 3). How closely the predicted performance agrees with that of real grape clusters will be the focus of future study. In addition, future work will explore the ability to detect real grapes within scans from 3D cameras using novel and existing algorithms. The accuracy of measurements of berry size distributions and grape bunch weight will then be investigated.

References

[1] F. Marinello, A. Pezzuolo, D. Cillis, L. Sartori, Kinect 3D reconstruction for quantification of grape bunches volume and mass, Engineering for Rural Development 15 (2016) 876–881.

[2] F. Marinello, A. Pezzuolo, F. Meggio, J. Martínez-Casasnovas, T. Yezekyan, L. Sartori, Application of the kinect sensor for three dimensional characterization of vine canopy, Advances in Animal Biosciences 8 (2) (2017) 525–529.

[3] D. Gao, Identification and location derivation of grapevine features through point clouds, Master's thesis, University of Adelaide (2014).

[4] F. Schöler, V. Steinhage, Automated 3D reconstruction of grape cluster architecture from sensor data for efficient phenotyping, Computers and Electronics in Agriculture 114 (2015) 163–177.

[5] Y. W. Kuan, N. O. Ee, L. S. Wei, Comparative study of Intel R200, Kinect v2, and Primesense RGB-D sensors performance outdoors, IEEE Sensors Journal 19 (19) (2019) 8741–8750.

[6] H. Sarbolandi, D. Lefloch, A. Kolb, Kinect range sensing: Structured-light versus time-of-flight Kinect, Computer Vision and Image Understanding 139 (2015) 1–20.

[7] K. Khoshelham, Accuracy analysis of Kinect depth data, in: ISPRS Workshop Laser Scanning, Vol. 38, 2011, p. W12.

[8] N. M. DiFilippo, M. K. Jouaneh, Characterization of different Microsoft Kinect sensor models, IEEE Sensors Journal 15 (8) (2015) 4554–4564.

[9] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, A. El Saddik, Evaluating and improving the depth accuracy of Kinect for Windows v2, IEEE Sensors Journal 15 (8) (2015) 4275–4285.

[10] E. Lachat, H. Macher, T. Landes, P. Grussenmeyer, Assessment and calibration of a RGB-D camera (Kinect v2 sensor) towards a potential use for close-range 3D modeling, Remote Sensing 7 (10) (2015) 13070–13097.

[11] S. Zennaro, M. Munaro, S. Milani, P. Zanuttigh, A. Bernardi, S. Ghidoni, E. Menegatti, Performance evaluation of the 1st and 2nd generation Kinect for multimedia applications, in: Multimedia and Expo (ICME), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.

[12] M. G. Diaz, F. Tombari, P. Rodriguez-Gonzalvez, D. Gonzalez-Aguilera, Analysis and evaluation between the first and the second generation of RGB-D sensors, IEEE Sensors journal 15 (11) (2015) 6507–6516.

[13] M. Carfagni, R. Furferi, L. Governi, M. Servi, F. Uccheddu, Y. Volpe, On the performance of the Intel SR300 depth camera: metrological and critical characterization, IEEE Sensors Journal 17 (14) (2017) 4508–4519.

[14] T. Mallick, P. P. Das, A. K. Majumdar, Characterizations of noise in Kinect depth images: A review, IEEE Sensors journal 14 (6) (2014) 1731–1740.

[15] I.-S. Kweon, J. Jung, J. Y. Lee, Noise aware depth denoising for a time-of-flight camera, in: 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, Asian Federation of Computer Vision (AFCV), 2014.

[16] T. Luhmann, K. Wendt, Recommendations for an acceptance and verification test of optical 3-D measurement systems, International Archives of Photogrammetry and Remote Sensing 33 (B5/2; PART 5) (2000) 493–500.

[17] J. Boehm, Accuracy investigation for structured-light based consumer 3D sensors, Photogrammetrie-Fernerkundung-Geoinformation 2014 (2) (2014) 117–127.

[18] K. Konolige, P. Mihelich. Technical description of Kinect calibration [online].

[19] M. Carfagni, R. Furferi, L. Governi, C. Santarelli, M. Servi, F. Uccheddu, Y. Volpe, Metrological and critical characterization of the Intel D415 stereo depth camera, Sensors 19 (3) (2019) 489.

[20] A. Myronenko, X. Song, Point set registration: Coherent point drift, IEEE transactions on pattern analysis and machine intelligence 32 (12) (2010) 2262–2275.

Chapter 3

Analysis of Depth Cameras for Proximal Sensing of Grapes

This chapter is republished in accordance with MDPI's copyright policy. The work presented here is the accepted version of the published article. Therefore, the contents are the same but there may be stylistic differences to the published article.

 B. Parr, M. Legg, F. Alam. Analysis of depth cameras for proximal sensing of grapes. Sensors (2022) 22(11), 4179. doi.org/10.3390/s22114179

Analysis of Depth Cameras for Proximal Sensing of Grapes

Baden Parr, Mathew Legg*, Fakhrul Alam

Department of Mechanical and Electrical Engineering, Massey University, Auckland, New Zealand

Abstract

This work investigates the performance of five depth cameras in relation to their potential for grape yield estimation. The technologies used by these cameras include structured light (Kinect V1), active infrared stereoscopy (RealSense D415), time of flight (Kinect V2 and Kinect Azure), and LiDAR (Intel L515). To evaluate their suitability for grape yield estimation, a range of factors were investigated including their performance in and out of direct sunlight, their ability to accurately measure the shape of the grapes, and their potential to facilitate counting and sizing of individual berries. The depth cameras' performance was benchmarked using high-resolution photogrammetry scans. All the cameras except the Kinect V1 were able to operate in direct sunlight. Indoors, the RealSense D415 camera provided the most accurate depth scans of grape bunches, with a 2 mm average depth error relative to photogrammetric scans. However, its performance was reduced in direct sunlight. The time of flight and LiDAR cameras provided depth scans of grapes that had about an 8 mm depth bias. Furthermore, the individual berries manifested in the scans as pointed shape distortions. This led to an underestimation of berry sizes when applying the RANSAC sphere fitting but may help with the detection of individual berries with more advanced algorithms. Applying an opaque coating to the surface of the grapes reduced the observed distance bias and shape distortion. This indicated that these are likely caused by the cameras' transmitted light experiencing diffused scattering within the grapes. More work is needed to investigate if this distortion can be used for enhanced measurement of grape properties such as ripeness and berry size.

Keywords: grapes, yield estimation, depth cameras, RGB-D

1. Introduction

Accurate and timely yield estimation can have a significant effect on the profitability of vineyards. Among other reasons, this can be due to better management of vineyard logistics, precise application of vine inputs, and the delineation of grape quality at harvest to optimise returns. Traditionally, the process of

*Corresponding author
Email addresses: 1badenparr@gmail.com (Baden Parr), M.Legg@massey.ac.nz (Mathew Legg), F.Alam@massey.ac.nz (Fakhrul Alam)

yield estimation is conducted manually. However, this is destructive, labour-intensive and time-consuming leading to low sampling rates and subjective estimations [1]. Automating yield estimation is therefore the focus of ongoing research in the computer vision field [2].

Current 2D camera techniques predominantly rely on distinct features of grapes, such as colour or texture, to identify and count individual berries within RGB (Red, Green, and Blue) images [3, 4]. However, the accuracy of yield estimations from these approaches is greatly restricted by the proportion of grapes visible to the camera. Hence, occlusion of grapes is an issue. Additionally, errors in the sizing of grapes can occur unless the distance between the camera and the grapes is known.

An alternative technique, which has been reported to provide improved yield accuracy, has been to incorporate 3D information. Grape bunch 3D architectonic modelling has been performed from high-resolution 3D scans of grape bunches within lab environments. These have been achieved using commercial laser scanners [5, 6] and blue LED structured light scanners [7, 8, 9, 10]. These scans can be used to estimate volume, mass, and number of berries per bunch. However, these 3D scanners are costly, require significant time to capture viable point clouds, and their use is yet to be demonstrated within field environments.

High-resolution 3D scans of grapes and vines have also been achieved using multiple RGB images captured from different positions using structure from motion photogrammetry techniques [11, 12, 13]. This method can be used with inexpensive equipment [14] and data collection can be automated by mounting cameras on platforms such as robots or drones [15]. However, generating photogrammetry scans requires significant computation load and time. Rose et al. [12] quoted 8 hours to generate a point cloud for one 25 m length of vine.

An alternative approach that has been investigated is to identify within an RGB image the location and size of individual berries within a bunch and use this information to model the 3D grape bunch architecture using spheres or ellipsoid shapes. Liu et al. [16, 17, 18, 19] used a backing board behind the grape bunch when capturing the RGB images to aid with the segmentation of individual berries. Berry size was estimated by placing a chequerboard pattern on the board. This allowed the distance between the camera and the backing board to be measured using camera calibration techniques. However, this requirement for a backing board means it can only be used for handheld applications. Ivorra et al. demonstrated/developed a novel technique that utilised a stereoscopic RGB-D (Red, Green, Blue—Depth) camera to obtain berry size without having to use a chequerboard pattern. They combined the depth information with 2D image analysis to achieve 3D modelling of the grape bunches.

The potential real-time benefits of RGB-D cameras for grape yield estimation have encouraged researchers to investigate their use for grape yield estimation. A range of low-cost RGB-D cameras that can generate 3D scans in real-time has become available on the market in recent years. This has been driven by their use in a wide range of applications including gaming, robotics, and agriculture. The main technologies used are stereoscopy, Active Infrared Stereoscopy (AIRS), Structured Light (SL), Time of Flight (ToF),

and Light Detection And Ranging (LiDAR). Stereoscopy is similar to human vision and uses parallax and disparity between featured in images from cameras that are spatially separated. Active infrared stereoscopy is similar but projects an Infrared (IR) pattern into the scene to assist with finding correspondences. This is particularly useful for cases where objects being scanned have low visible texture and/or are in low light conditions. Structured light detects distortions in a known projected IR pattern. Time of flight and LiDAR cameras both operate by measuring the time taken for emitted IR light to be reflected back to the camera. ToF cameras typically emit this light in a single pulse, while LiDARs typically measure by sweeping a laser. RGB-D cameras have been used for 3D imaging a range of different fruits [20]. This includes several studies related to imaging grapes.

Marinello et al. [21] used a Kinect Version 1 (V1) camera, which operates using IR structured light, to image grapes in a lab environment for yield estimation. Their results showed that the scanning resolution decreased significantly with the increased distance of the sensor from the grapes. Hacking et al. [22, 23] also used the Kinect V1 for yield estimation in both lab and vineyard environments. They showed that the Kinect V1 gave a good correlation with grape bunch volume in the lab but struggle in the field environment. They suggested that this could be due to sunlight and the presence of leaves. They recommended that future work should investigate the performance of the Kinect V2, since it is a ToF camera and hence is more robust to sunlight conditions compared with SL cameras, such as the Kinect V1, which project IR patterns [24]. An alternative approach could be to take measurements at night. This technique has been used by studies capturing traditional RGB images in vineyards [3, 25].

Kurtser et al. [26] used an Intel RealSense D435 RGB-D camera, which operates using AIRS technology, for imaging grapes bunches in an outdoor environment. They used neural networks for detecting grape bunches from the point clouds [27]. Basic shapes (box, ellipsoid, and cylinder) were fitted to the point clouds. However, they reported relatively large (28–35 mm) errors in the length and width of these fitted shapes compared with the physical measurement of the grape bunches. These errors were reported to be affected by sunlight exposure. It would appear that in sunlight conditions, the projected IR pattern would not be viable meaning this camera would be acting as a stereo camera.

Ivorra et al. [28] used a stereoscopic RGB-D camera (Point Grey Bumblebee2) for imaging grapes, as mentioned above. However, the 3D scans of the grapes from this camera were of poor quality. They suggested that this was due to difficulty in making the correct correspondence between the stereo image pairs. Yin et al. [29] also used a stereoscopic camera (ZED) for imaging grapes. However, this was used to measure the pose of grape bunches for automated robotic picking rather than yield estimation.

This article presents the first benchmarking of the performance of multiple RGB-D cameras for use in grape yield estimation applications. This includes ToF cameras, which have not been used before in a grape yield estimation study. The benchmarking performance analysis was obtained by calculating error maps between high-resolution scans obtained using photogrammetry and those obtained by the RGB-D cameras.

This includes an analysis of the cameras' performance in and out of direct sunlight.

Previous studies [21, 22, 23, 26, 27, 28] have only looked at volume errors for a grape bunch as a whole. However, in this work, depth map errors in the RGB-D scans of grapes are analysed at an individual grape berry scale, which has not been done before.

The ability to identify individual grapes from 3D scans would provide additional information for the yield and crop load estimation process. This could inform viticulturists of metrics such as berry size distribution and berry count per cluster. There is also the potential for more accurate volume estimates by 3D modelling of the grape cluster architecture. This has been explored by several researchers [5, 6, 7, 8, 9, 10, 16, 17, 18, 19, 28] but not for RGB-D cameras. This might be because it has been thought that these cameras did not have sufficient precision [5].

In this work, the ability of RGB-D cameras for detecting individual grape berries using Random Sample Consensus (RANSAC) is investigated. We are not aware of any reported works that have applied an algorithm such as RANSC with RGB-D camera scans for grape berry detection.

The remainder of the article is organised as follows. Section 2 describes the experimental setup and data processing used. The results are presented in Section 3. Section 4 provides a discussion on the results. Finally, a conclusion is provided in Section 5 .

2. Methodology

2.1. Hardware and Measurement Procedure

The RGB-D cameras used in this work were chosen to cover the main technologies available. The cameras used were the Kinect V1 (SL), Intel RealSense D415 (AIRS), Microsoft Kinect V2 (ToF), Microsoft Kinect Azure (ToF), and Intel L515 (LiDAR). Table 1 provides some specifications on these cameras. Additionally, a Sony Alpha A6300 mirrorless RGB camera was used to obtain high-resolution scans of the grapes using photogrammetry. Note that the Kinect V1 and Kinect V2 are discontinued. However, the Kinect V2 is still very commonly used in research and both are used or mentioned in the related literature. Including the results from these two cameras also provides benchmarking of the newer with older camera technologies.

The RGB-D cameras were mounted on a 2D gantry (CNC machine). The gantry had a 2D travel range of 1400 $\times$ 1400 mm and a resolution of 0.025 mm. A bunch of green table grapes was suspended in front of the cameras at one end of the gantry. The gantry system was used to move the camera under investigation directly in front of and at the desired distance from the grapes, see Figure 1.

Table 1: List of RGB-D cameras used with the depth measurement technologies they use and their resolution and field of view specifications.

Camera	Technology	Resolution [Pixels]	Field of View [Deg]
RealSense D415	AIRS	1280 × 720	65 × 40
Kinect V1	SL	640 × 480	57 × 43
Kinect V2	ToF	512 × 424	70 × 60
Kinect Azure	ToF	1024 × 1024	120 × 120
Intel L515	LiDAR	1024 × 768	70 × 55

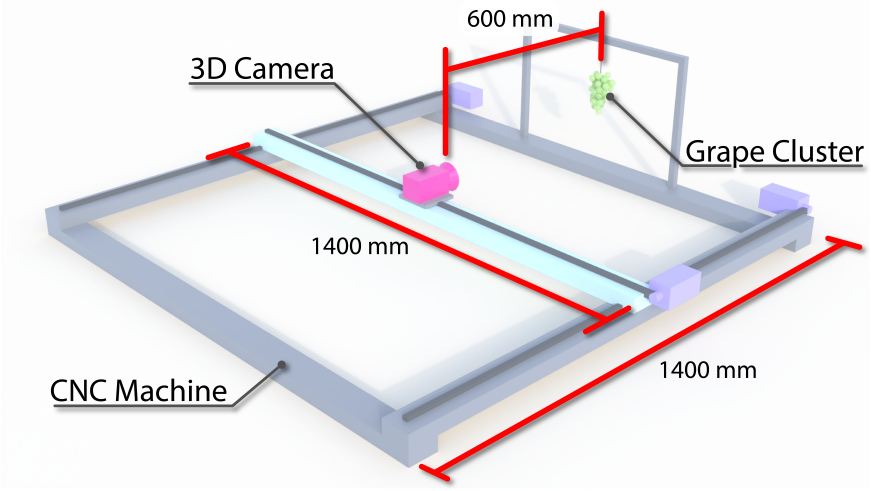
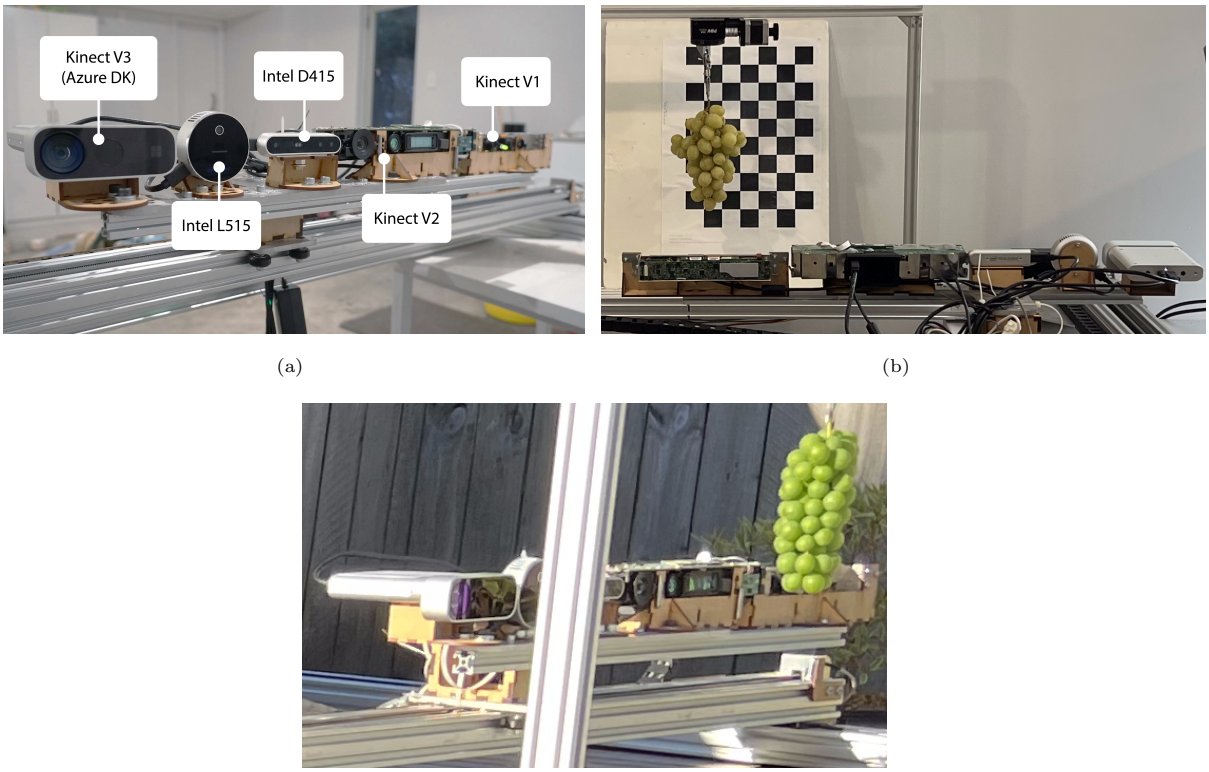


Figure 1: Diagram showing setup of camera and grapes mounted onto the CNC machine for capturing RGB-D images.

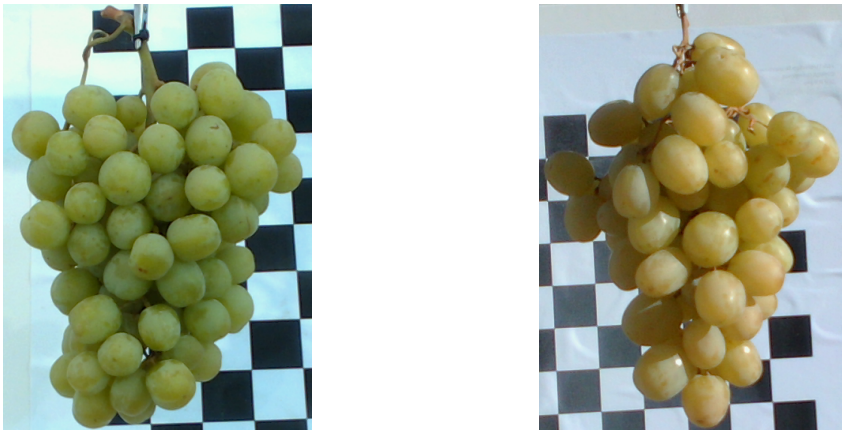
Figure 2 provides photos of the experimental setup. Figure 3 shows photos of the grapes used in this work for both indoor and outdoor measurements. These are cropped versions of the images captured by the Intel L515 camera, which was located 600 mm from the grapes.



(a) (b)

(c)

Figure 2: Photos of the experimental setup. Photo (**a**) shows the front view of the cameras mounted onto the 2D gantry. Photos (**b**,**c**) respectively show the setup located inside and outdoors.



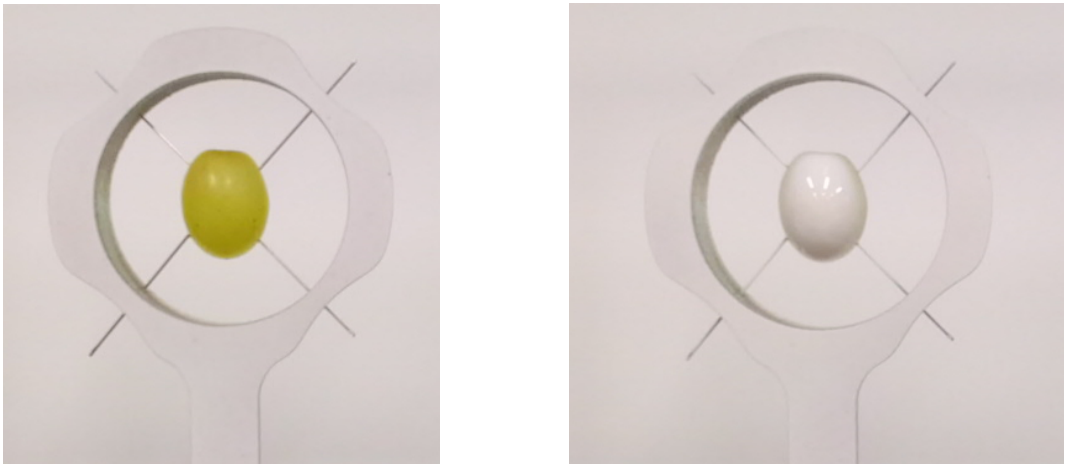
(a) (b)

Figure 3: Coloured images of the grape bunches used in this work for scans captured (**a**) indoors and (**b**) outdoors.

Python code was used to move the gantry so that a camera under investigation was directly in front

of the grape bunch and then capture RGB-D images with the camera at a number of distances from the grapes. Most of the measurements shown in this work were with the camera located at a distance of 600 mm from the grapes. This distance was used as it was a distance that worked well for all cameras tested. For example, the Kinect V1 and V2 struggled to capture images at distances much closer than this. The newer cameras were able to image at closer ranges. In addition to this, it was felt that this distance was likely to be a practical separation distance of the cameras from the grapes if the camera was mounted onto a farm vehicle driving between vine rows. This process was then repeated for all the RGB-D cameras. The Sony Alpha A6300 mirrorless RGB camera was then used to capture RGB (6000 $\times$ 4000 pixel) images of the grapes at a range of positions for high-resolution photogrammetry scans. The above measurement process was performed first in the lab and then outdoors in direct sunlight using a different grape bunch. This was done to evaluate the effect of sunlight on the performance of each RGB-D camera.

Measurements were also performed to evaluate if diffused scattering within the grapes was causing distortions in the ToF and LiDAR cameras. This was achieved by obtaining scans before and after spraying the grapes with white primer paint. The paint aimed to make the grapes opaque and hence stop diffused scattering within the berries. Figure 4 shows the setup used for a single grape positioned inside a ring before and after it has been sprayed with paint. Needles were used to secure the grape and ensure that the front face of the grape was flush with the front surface of the ring. Care was taken to not pierce the grape so as not to disrupt the internal optics of the grape.



(a) (b)

Figure 4: Photos of the setup of scans for a single grape which is first unpainted (**a**) and then painted (**b**). This was performed to analyse the effect of diffused scattering within the grape for the RGB-D cameras, which use ToF and LiDAR technologies.

2.2. Processing Data

The software Agisoft Metashape v1.5.2 was used to obtain high-resolution photogrammetry scans of the grape bunches using the RGB images captured by the Sony A6300 from a number of positions. These provided a baseline scan that could be used to evaluate the accuracy of the RGB-D cameras. The point clouds obtained using both the RGB-D and photogrammetry scans were then processed using CloudCompare. This is a widely used 3D point cloud and mesh processing open-source software. It has a range of point cloud processing tools including cropping, filtering, alignment, distance measurement, and comparison of multiple point clouds.

It was observed that the raw ToF and LiDAR camera scans had a significant number of flying pixels around the edges of the grape bunch. A significant portion of these was therefore filtered out using CloudCompare. This was done by rejecting points that had normal angles greater than a set value. This was empirically chosen to be 85 degrees. Isolated points were then discarded using statistical outlier rejection, which compared distances between its six nearest neighbours and used one standard deviation of the entire point cloud distribution as the rejection threshold [30].

2.2.1. Alignment of Scans and Generating Error Maps

The RGB-D camera scans needed to be aligned with the photogrammetry scan in order to allow benchmarking to be performed. Let $\boldsymbol{X}_i$ be a $[3 \times N]$ coordinate vector of the N selected points on the RGB-D scan and $\boldsymbol{Y}_j$ be the corresponding coordinates of the selected points in the photogrammetry scan. Alignment of the RGB-D scan scans can then be achieved by finding the $[3 \times 3]$ rotation matrix $\boldsymbol{R}$ and the $[3 \times 1]$ translation vector $\boldsymbol{T}$ such that when the RGB-D scan undergoes a ridged body translation the distance between the selected RGB-D and photogrammetry scan points are minimised. This can be expressed as

$$[\boldsymbol{R}^*, \boldsymbol{T}^*] = \underset{\boldsymbol{R},\boldsymbol{T}}{argmin} \sum_{i,j} ||\boldsymbol{Y}_j - \boldsymbol{R}\ \boldsymbol{X}_i - \boldsymbol{T}||^2. \tag{1}$$

Rather than aligning the two scans using manually selected points, the alignment can also be performed automatically using cropped RGB-D and photogrammetry scans and solving Equation (1) using a process referred to as the Iterative Closest Point (ICP) algorithm. Refer to Zinßer et al. [31] for more details on the ICP algorithm used by CloudCompare [32]. The optimised values of $\boldsymbol{R}$ and $\boldsymbol{T}$ can then be used to perform the ridged body translation

$$\overline{\boldsymbol{X}} = \boldsymbol{R}^*\ \boldsymbol{X} + \boldsymbol{T}^*. \tag{2}$$

on the RGB-D scan to align it with the photogrammetry scan.

The alignment process described above was initially performed using CloudCompare and manual selection of points on the chequerboard image for both scans. The point clouds were then cropped to just include the grape bunch. An error scan for each RGB-D camera was then obtained. This was calculated by measuring

the distance from each point in an RGB-D camera's scan to the closest point in the photogrammetry scan [33]. Refer to Figure 5 for a block diagram summarising the processing steps used to obtain the depth error maps.

An alternative error analysis method was also used, which aligned the depth camera and photogrammetry scans of the grape bunch using the ICP algorithm, rather than using the chequerboard image. The raw scans were cropped in CloudCompare to just include the scans of the grape bunch. Scaling was also performed on the RGB-D camera scan to correct for projection if this scan was located behind the photogrammetry scan, due to any diffused scattering within the berries. Alignment between the RGB-D scan and the photogrammetry scan was performed using an ICP algorithm. The error in the RGB-D scan was obtained by finding the distance from each point in the ICP aligned RGB-D scan to the closest point in the photogrammetry scan.

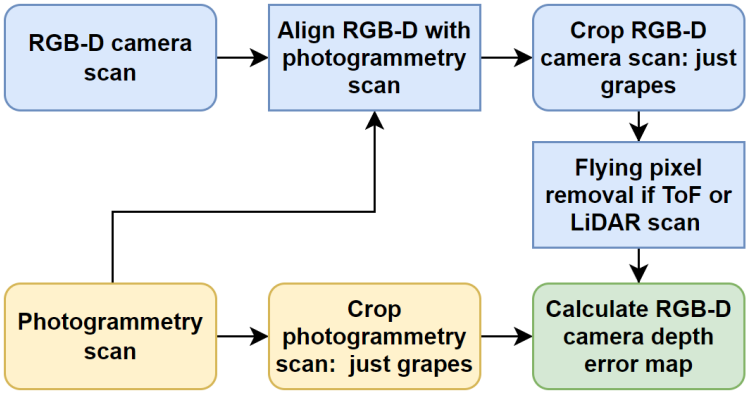


Figure 5: Diagram showing the processing steps used to calculate error depth maps for the RGB-D cameras.

2.2.2. Calculating the Proportion of Missing Scan Points

Image processing was performed to estimate the proportion of the scan that was missing for each depth camera relative to the photogrammetry scan. CloudCompare was used to capture 2D images of each depth camera's scan of the grapes with a white background. To ensure consistency between cameras, these images were obtained using the same viewing angle and position and image size. The percentage of pixels in this image that was white (not grapes) was then calculated using MatLab for each depth camera. The percentage of missing scan area was then obtained by subtracting this value from that obtained for the photogrammetry scan.

2.2.3. Identifying Individual Grapes Using RANSAC

Work was also performed to investigate the potential of identifying and sizing individual grapes from the RGB-D camera scans. The RANSAC algorithm was chosen as it is the method that has been used in the literature related to identifying the position and size of grapes from high-resolution scans captured using

commercial scanners. This algorithm fits shapes such as spheres to the scan. Ideally, the size and position of each grape can be identified from the size and position of the corresponding fitted sphere.

CloudCompare was used to apply the RANSAC algorithm to the indoor scans obtained using both the RGB-D cameras and photogrammetry. Schnabel et al. [34] provides a description of the RANSAC algorithm used by CloudCompare [35]. It fitted spheres to the grape bunch scans and used this to segment the scans into a single point cloud for each fitted sphere. Ideally, each of these segmented point clouds would correspond to a different grape. These point clouds were then exported as separate files with the sphere radius in the file name. However, it did not contain the location of the sphere's centroid.

MatLab was then used to process these segmented point clouds using the least-squares sphere fitting function provided in [36]. For a given camera, each file was loaded and a least-squares fitting of a sphere to the segmented scan was performed to obtain the position of the sphere's centroid. The closest sphere in the photogrammetry scan was then identified using a K-Nearest Neighbours (KNN) search.

The difference in the 2D position of the RGB-D camera's sphere relative to the corresponding sphere for the photogrammetry scan was then calculated. This distance calculation did not include an offset in the depth axis direction. This was done to avoid this measurement being dominated by any distance bias that might be present for the depth cameras. Similarly, the difference in the RGB-D camera's fitted sphere radius relative to the corresponding photogrammetry sphere was also calculated. This process was repeated for all the segmented point clouds and median values obtained. Note that the median was used rather than the average since several fitted spheres were too large relative to the size of the grapes and would have distorted the averaged results. Spheres with a radius greater than 20 mm were ignored when counting the number of fitted spheres.

3. Results

Photogrammetry point clouds of the grape bunches were obtained to act as baseline scans which could be used to evaluate the accuracy of the RGB-D camera scans. Figure 6 provides an example of a high-resolution scan obtained using photogrammetry of the grape bunch for the indoor scans. This scan was obtained using RGB images captured by the Sony A6300 camera. Note that the depth colour scale is relative to the minimum and maximum depth value and has been normalised so that the closest point on the grapes is set to 0 mm. This allows comparisons of depth maps to be made across cameras.

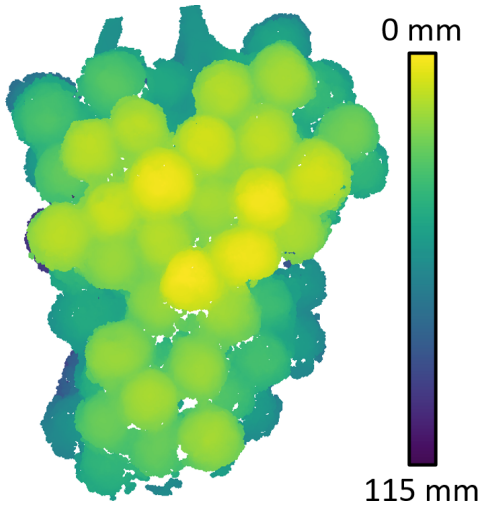
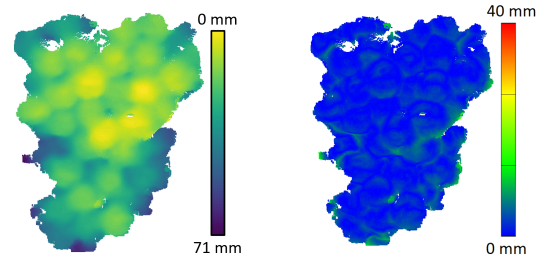


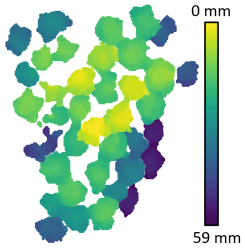
Figure 6: Example photogrammetry 3D depth scan of the grape bunch which was located indoors.

The photogrammetry scan was used as a ground truth to obtain error maps for depth scans captured by the RGB-D cameras. Figure 7 shows the depth and error scans of the RGB-D cameras, which were captured indoors with the cameras located at a distance of 600 mm from the grapes. Note that these error scans were obtained by aligning the depth camera and photogrammetry scans using the chequerboard image and not using the ICP alignment method. These results show that the ToF and LiDAR cameras give depth scans of the grape clusters that had distances biased to be further away than they should be. This effect was not observed for the Kinect V1 or the RealSense D415 cameras. It was believed that diffused scattering in the grapes could be the cause of the distance bias for the ToF and LiDAR cameras. The following section investigates this further.

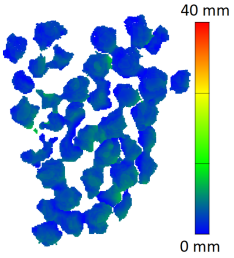


(a) RealSense D415 depth.

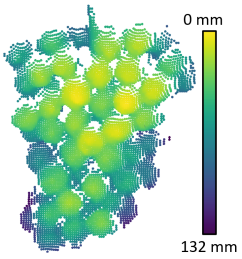
(b) RealSense D415 error.



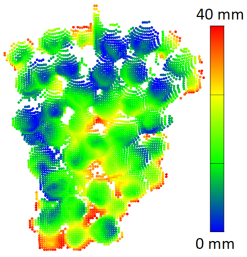
(c) Kinect V1 depth.



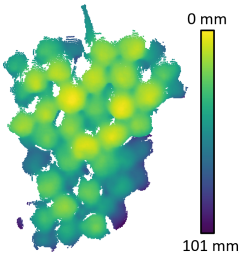
(d) Kinect V1 error.



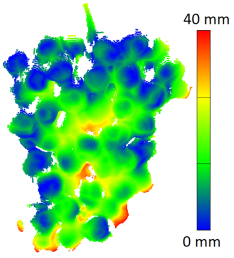
(e) Kinect V2 depth.



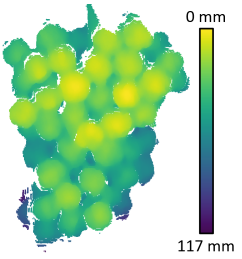
(f) Kinect V2 error.



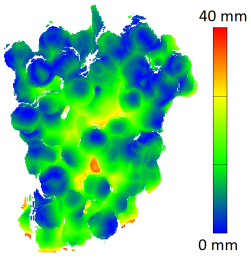
(g) Kinect Azure depth.



(h) Kinect Azure error.



(i) Intel L515 depth.

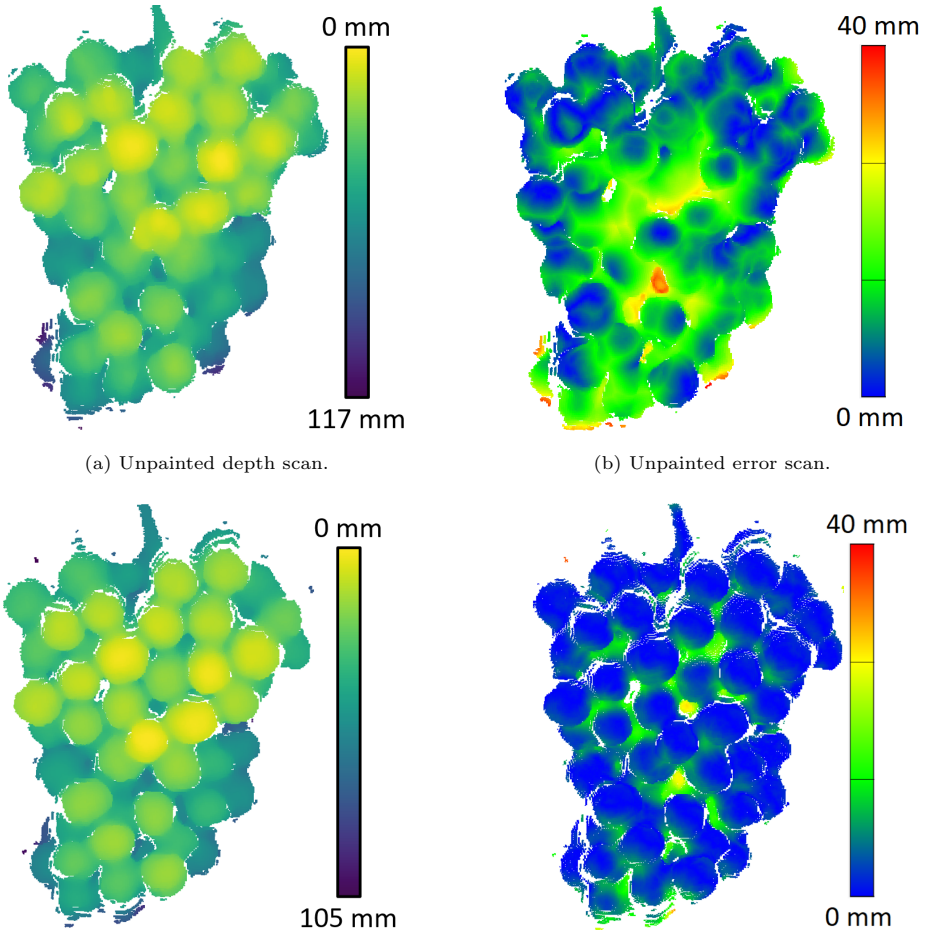


(j) Intel L515 error.

Figure 7: Depth and error scans (relative to the photogrammetry scans) for the RGB-D cameras located indoors at a distance of 600 mm from the unpainted grape bunch. An error bar is provided that shows the colour scale for the error scans and is the same for all the cameras. The colours for the depth scans are relative to the maximum and minimum depth of the point cloud for each camera.

3.1. Investigation of Distortion Effects

The grapes were spray-painted with white paint to investigate if diffused scattering was causing the distance bias for ToF and LiDAR cameras. Figure 8 provides examples of the Intel L515 LiDAR depth scans for a grape bunch before and after it had been sprayed with paint. The painted scans have the depth error bias removed and the clarity of individual berries in the depth map appears to be slightly enhanced.



(a) Unpainted depth scan.

(b) Unpainted error scan.

(c) Painted depth scan.

(d) Painted error scan.

Figure 8: Depth and error scans for the Intel L515 before and after spray painting the grape bunch with white paint.

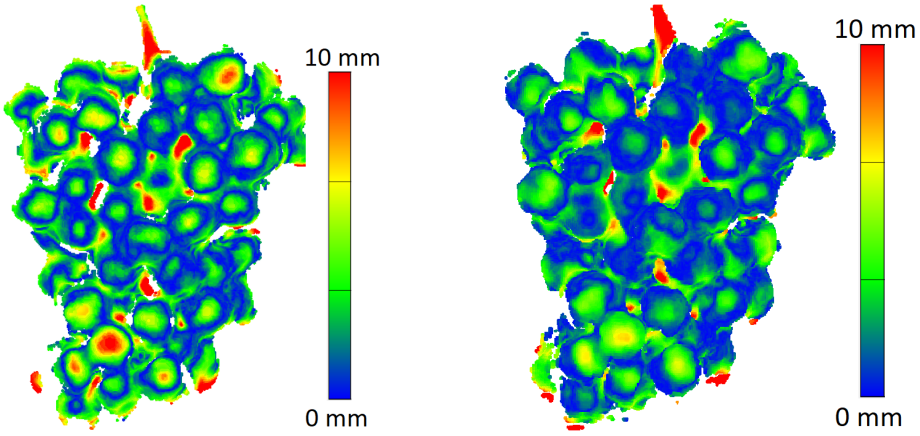
Table 2 provides the mean distance error for the grapes bunch for scans made before and after the grapes were spray-painted. No significant difference in the error (only 0.5 mm) was observed between the unpainted and painted scans for the Kinect V1 and RealSense D415, which are SL and AIRS cameras. However, we can see that painting the grapes reduces the distance bias for the ToF and LiDAR cameras.

Table 2: Mean depth error for RGB-D camera scans of the grapes before and after they had been sprayed with paint. The cameras were located indoors and were positioned 600 mm from the grapes.

Camera	Unpainted [mm]	Painted [mm]	Unpainted with ICP Alignment [mm]
RealSense D415	2.13	1.88	1.33
Kinect V1	3.67	3.00	1.01
Kinect V2	14.7	8.28	4.73
Kinect Azure	11.9	4.19	2.66
Intel L515	10.0	3.82	2.17

ICP alignment error analysis was also performed. This method appears able to remove the distance bias in post-processing, see the third column of Table 2. However, the errors for the ToF and LIDAR cameras are still slightly higher than their SL and AIRS counterparts.

Figure 9 shows the error maps for the Kinect Azure and Intel L515 cameras where ICP had been used to align their RGB-D depth scans with the photogrammetry scans. While this appears to have removed the distance bias, it shows that shape distortion errors still occur in the form of peaks located at the centre of each individual grape. The ToF cameras appeared to show slightly more pronounced shape distortions compared to the LiDAR.



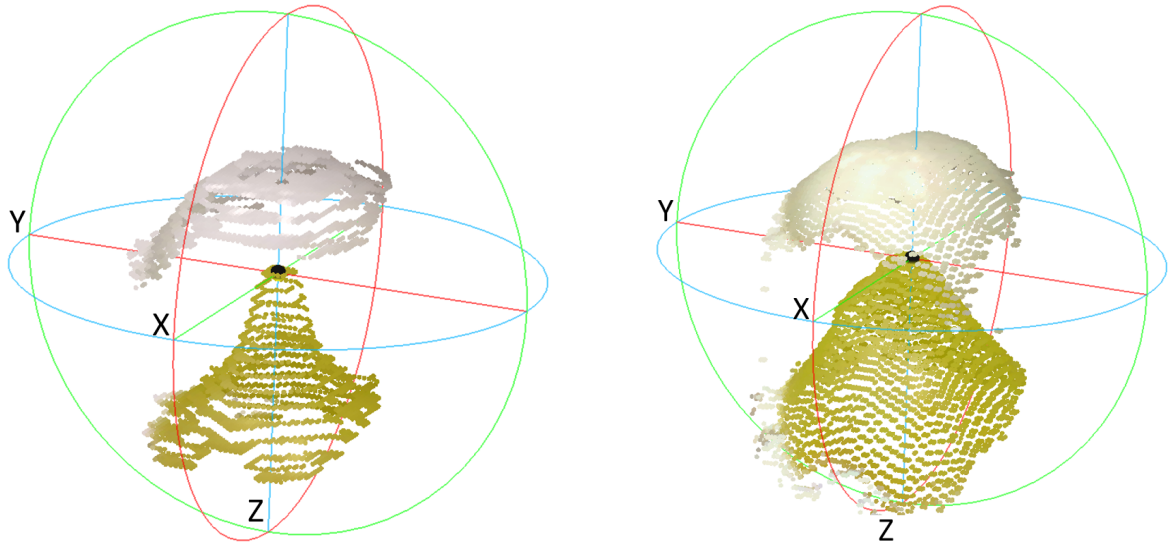
(a) Kinect Azure.

(b) Intel L515.

Figure 9: Kinect Azure and Intel L515 error scans for unpainted grapes after the depth camera scans were aligned with the photogrammetry reference scan using the ICP alignment method.

This distortion effect is illustrated in Figure 10. This plot shows scans captured by the Kinect Azure and Intel L515 of this grape before and after it was painted. These RGB-D cameras were located at a distance of 350 mm from the grapes. This distance was chosen as the distortion appeared slightly more pronounced

at this distance, as is illustrated in Figure 11. The unpainted grape scan points show significantly more pointed shape distortion compared with the painted grape.



(a) Kinect Azure ToF camera.

(b) Intel L515 LiDAR.

Figure 10: Scans for Kinect Azure (**a**) and Intel L515 (**b**) RGB-D depth scans of a single grape before (green) and after (white) individual grapes had been painted with white spray paint. Note that the Z-axis direction shown in the plots is the depth axis. The cameras were located 350 mm from the grapes. The Kinect Azure and Intel L515 have their unpainted peaks respectively about 7 mm and 8.5 mm behind the painted peaks. The Azure scan is more heavily quantised than the L515 scan.

Figure 11 shows cross-sections in the $X - Z$ plane of Kinect Azure scans made of a single grape before and after it had been sprayed with paint, for a range of distances of the depth camera from the grape. The depth has been normalised so that zero depth corresponds to the front of the ring supporting the grape. The distance bias and shape distortion are reduced when the grape is painted. It appears that the shape distortion is more pronounced when the camera is closer to the grape.

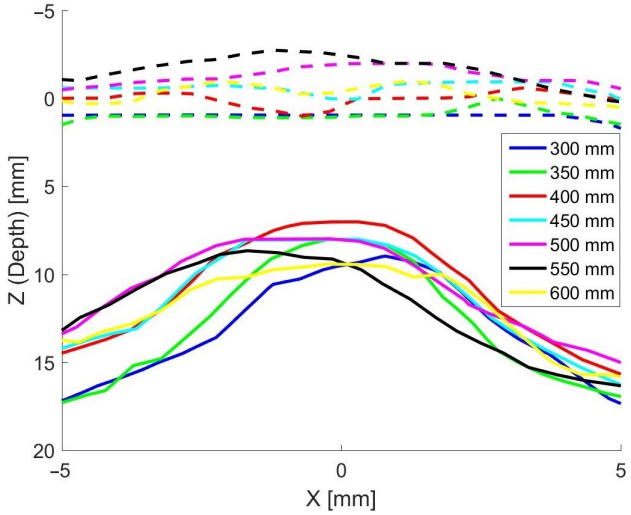
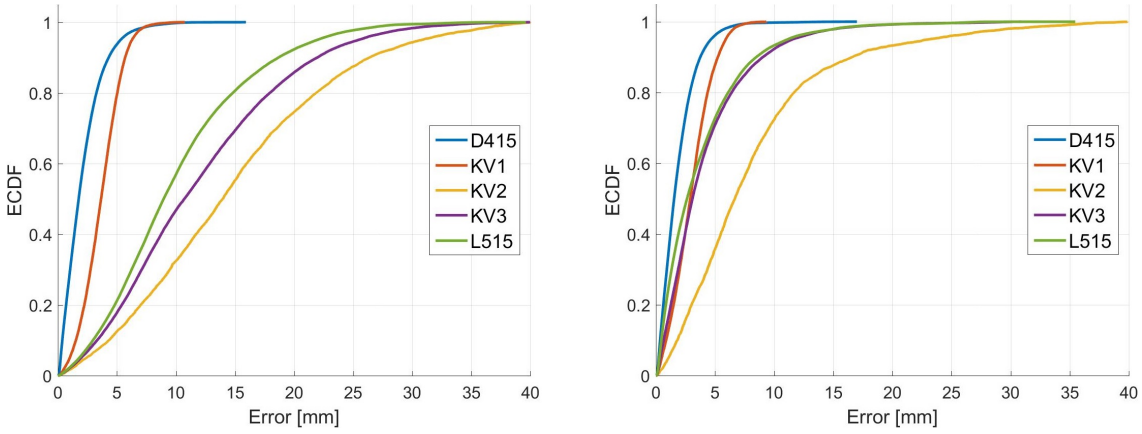


Figure 11: Plots showing cross-sections of scans made by the Kinect Azure of a single grape before (solid lines) and after (dashed lines) the grape had been painted. The different colours represent scans made with the camera being located at distances from the grape ranging from 350 to 600 mm.

Figure 12 provides plots of the Empirical Cumulative Distribution Functions (ECDF) of the errors in scans captured indoors both before and after the grape bunch had been sprayed with white paint. The ECDF plots show what percentage of the errors is below a given value. For example, we can see that, for the unpainted grapes, the Kinect V2 has 95% of its errors below 30 mm. In contrast, the corresponding scans for the RealSense D415 has 95% of its errors less than about 5 mm.



(a) Unpainted.

(b) Painted.

Figure 12: Plots (**a**,**b**) respectively show the ECDF error measurements for the grape bunch scans made indoors before and after the grapes had been sprayed with paint. The cameras were positioned 600 mm from the grapes.

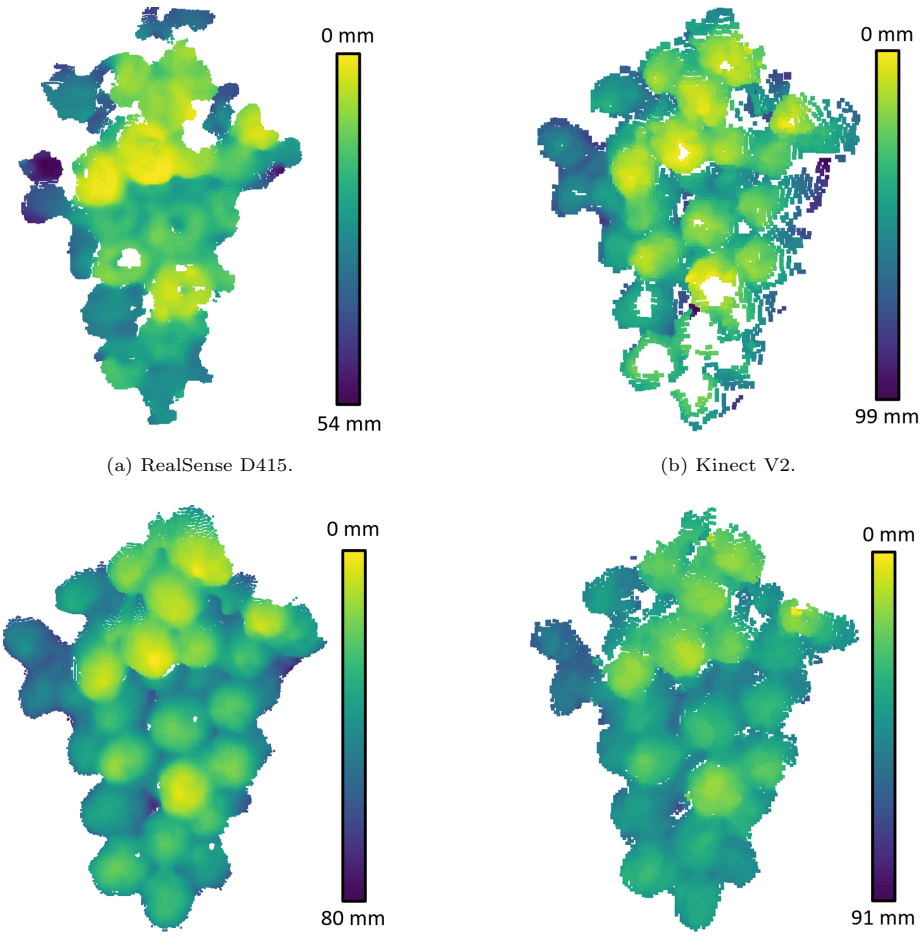
Note that some caution is required when interpreting the ECDF plots. This error analysis only looks

at errors in scan points captured with the depth cameras. However, it does not analyse how much of the scan was missing. For example, the ECDF plot shown in Figure 12 indicates that the Kinect V1 produced relatively low errors. However, from Figure 7 we can see that there was a significant proportion (about 20%) of the scan that was missing compared with the other cameras. Additionally, the ECDF does not provide information on how well individual grapes can be identified within a scan.

3.2. Measurements Made in Direct Sunlight

Measurements were also made using the cameras located outdoors to evaluate their performance in direct sunlight. Note that the grapes used for the indoor scans had been painted in order to investigate how diffused scattering within the berries affected the results. Hence, a different grape bunch was used for the outdoor scans. However, the methodology was designed with the aim of providing results that were independent of which grape bunch was used in the benchmarking by comparing the photogrammetry and RGB-D camera scans. This means that the error analysis should be relatively independent of the grape bunch used, though some difference in the results may occur.

Figure 13 shows examples of these depth scans with the cameras at a distance of 600 mm from the grapes. Note that no results are shown here for the Kinect V1. This is because no measurements were able to be achieved with this camera until after sunset. All of the other depth cameras were able to obtain scans of the grapes in direct sunlight. However, the errors for the RealSense 415 are similar to those of the Kinect V3 and LiDAR for outdoor measurements but are still lower than those for the Kinect V2.



(a) RealSense D415.

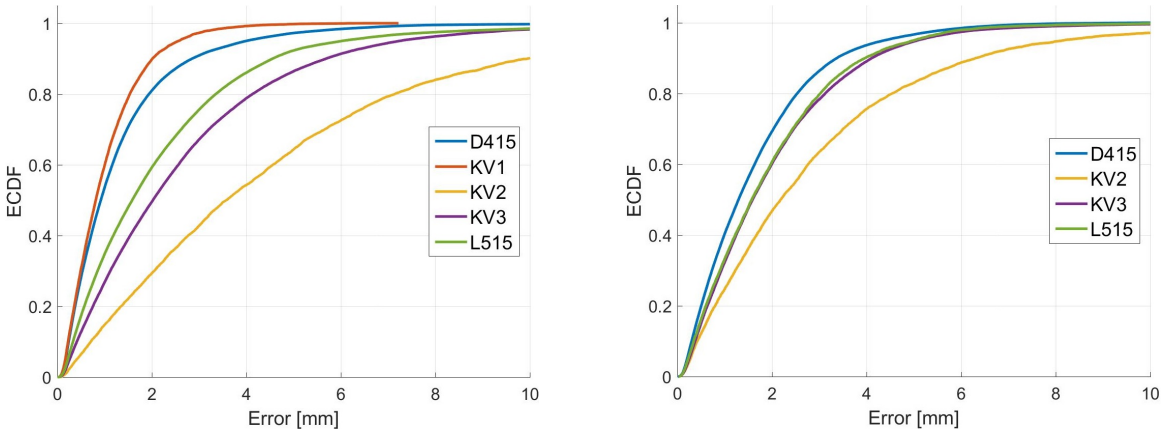
(b) Kinect V2.

(c) Kinect Azure.

(d) Intel L515.

Figure 13: Depth scans for the RGB-D cameras captured outdoors at a distance of 600 mm from the grape bunch.

Figure 14 compares ECDF plots for these scans made outdoors with the scans made indoors where ICP alignment has been used. Table 3 provides a comparison of the proportion of missing scan points for each camera for both indoor and outdoor measurements. It can be seen that the RealSense D415 has a 13% increase in the proportion of missing scan points for outdoor measurements, while the ToF and LiDAR cameras are relatively unaffected. There is a slight (2%) reduction in the proportion of missing scan points for the Kinect V2 outdoors relative to indoors. However, this is probably within the measurement error for this analysis method or may be due to the fact that different grape bunches were used for the indoor and outdoor experiments.



(a) Indoors.

(b) Outdoors.

Figure 14: Plots comparing ECDF plots for scans of two different unpainted grape bunches which were captured by depth cameras (**a**) indoors and (**b**) outdoors in direct sunlight using ICP alignment of the depth camera scans with the photogrammetry scans. The grapes were located 600 mm from the cameras.

Table 3: Estimate of the percentage of the depth scan that is missing for each camera relative to that obtained using the photogrammetry scans.

Camera	**Indoors [%]**	**Outdoors [%]**
RealSense D415	0.9	14
Kinect V1	20	–
Kinect V2	14	12
Kinect Azure	4.2	4.6
Intel L515	2.0	3.6

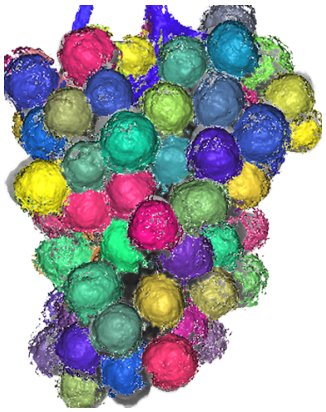
3.3. Detection of Individual Grapes Using RANSAC

Analysis was performed on the grape scans that were captured indoors to investigate if it was possible to detect and size individual grapes from the raw RGB-D camera depth scans. The RANSAC algorithm within CloudCompare was used to fit spheres to the depth scans. Figure 15 shows the resulting segmentation of the scans provided by the RANSAC sphere fitting for the photogrammetry and depth camera scans. These are overlaid over a photo of the grapes for comparison. The different colours correspond to different segmented point clouds obtained by fitting spheres to the raw scans. Ideally, there would be a separate colour for each grape. However, it can be seen that the results are not perfect. The performance of the algorithm is lower for the RGB-D cameras scans compared to that of the photogrammetry scan.

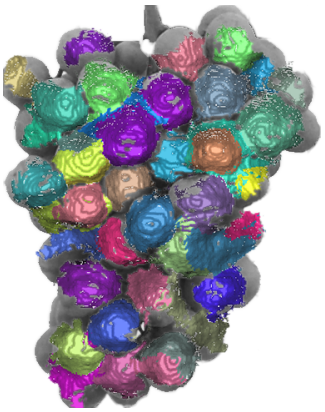
Table 4 provides the median difference in the detected 2D position and sphere radius relative to the corresponding spheres for the photogrammetry scans. The medium sphere radius for the photogrammetry

scans was 13.7 mm. The depth information was ignored when calculating the 2D position error since adding depth would have resulted in values that were dominated by the distance bias for the ToF and LiDAR cameras. The median differences in the 2D positions of the spheres are relatively low. These position errors may be related to errors in the alignment of the depth camera scans in comparison to the photogrammetry scan.

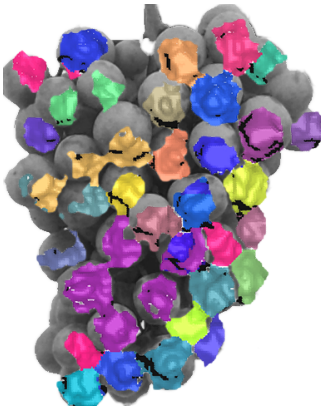
This table also gives the number of spheres detected for each RGB-D camera that had radius values less than 20 mm. We can also see that the ToF and LiDAR camera scans have smaller median sphere radius values compared to those obtained using photogrammetry and the RealSense D415 and Kinect V1 cameras.



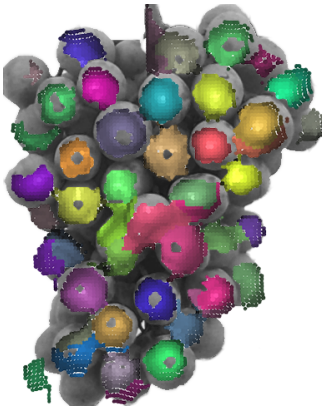
(a) Photogrammetry.



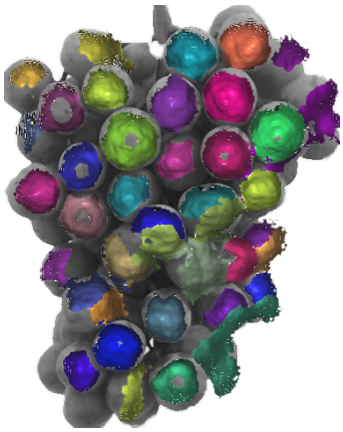
(b) RealSense D415.



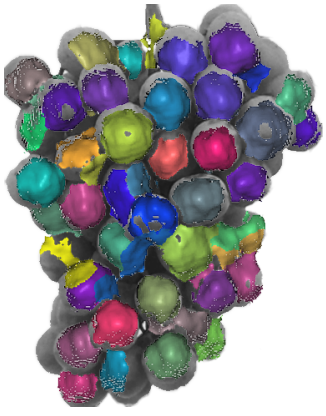
(c) Kinect V1.



(d) Kinect V2



(e) Kinect Azure.



(f) Intel L515.

Figure 15: Plots showing the output of the RANSAC algorithm on the grape depth scans which were captured indoors. This is overlaid over a greyscale photo of the grape bunch for reference.

Table 4: Information on the RANSAC algorithm fitting of spheres to individual grapes in the scans. This shows the number of spheres detected and the median difference in the radius and 2D positions of the spheres for the RGB-D cameras relative to the same spheres in the photogrammetry scans.

	RealSense D415	**Kinect V1**	**Kinect V2**	**Kinect Azure**	**Intel L515**
No. of Spheres Detected	25	22	31	26	30
Median Radius Difference [mm]	1.7	2.8	−3.7	−3.5	−3.0
Median Position Difference [mm]	1.9	2.1	3.5	2.0	2.1

4. Discussion

The RealSense D415, which uses AIRS technology, was the most accurate camera indoors. However, it showed reduced performance outdoors. This is in line with the findings of Kurtser et. al. [26] that reported increased errors for the RealSense D435 AIRS camera with increased sunlight exposure. The ECDF plots shown in Figure 14 indicate that the errors for the RealSense D415 increased outdoors but were still similar to that of the Kinect Azure and Intel L515 (after correcting for their distance bias using ICP). However, the RealSense D415 also had a significant increase in missing scan points when operated in direct sunlight. This is illustrated in Table 3, where the percentage of missing scan points relative to the photogrammetry scan increased from about 1% to 14% when measurements were made outdoors. Additionally, the 3D shape of individual grapes was less pronounced, which would make it harder to identify and measure the size of the grapes. This might be because it was not able to use its projected IR pattern due to saturation by sunlight. Saturation of the stereo IR cameras may also have occurred. Moreover, the camera may have struggled with the dynamic range caused by direct illumination from the sun with shadows.

The Kinect V1 SL camera also had low depth errors for measurements made indoors. However, Table 3 shows that it had about 20% of the scan points missing, which was the highest of any of the other cameras. This resulted in a smooth shaped scan of the grape bunch and did not display the valleys between grapes. This phenomenon can be seen in the plots presented by Marinello et al. and Hacking et al. [21, 22, 23]. The Kinect V1 has a significant deterioration in resolution as the distance of the grapes from the camera increases, as reported by Marinello et al. [21]. This appears to be related to the strong depth quantisation dependence on scan depth for this camera.

The Kinect V1 could not be used for scanning grapes outdoors in direct sunlight. This was expected since its projected IR pattern would have been saturated by the sunlight. Hacking et al. [22, 23] had also reported issues with its performance when used outdoors. They had therefore suggested that the Kinect V2 should be investigated for outdoor grape bunch scanning since it would be more robust to sunlight.

The cameras that used ToF technologies were found to be more robust to sunlight conditions. Both the Kinect Azure and Intel L515 appeared to provide similar results indoors and outdoors in direct sunlight.

The Kinect V2 had higher errors than the Azure and Intel L515. It was able to operate in sunlight but did have some issues with saturation resulting in scan points being missing. This may be addressed by adjusting the exposure in software.

The ToF and LiDAR cameras produced scans of the grapes that had a distance bias of about 8 mm and had a distortion in the shape of the scans of the grapes, which was not observed for the SL and AIRS cameras. The shape distortion for the ToF and LiDAR cameras makes individual grapes within the scan more prominent and easier to identify than the Kinect V1 and the RealSense D415. This distortion may therefore be beneficial for counting individual grapes. The plots in Figures 8, 10, and 12 show that these distortion effects were largely removed when the grapes were painted. This indicates that the distance bias and shape distortions are due to diffused scattering within the berries of the transmitted light used by these cameras.

The Intel L515 LiDAR appeared to have slightly less distance bias and distortion compared to the two ToF cameras. The difference in distortion between the ToF and LiDAR cameras may be due to the process they used to emit light. ToF cameras emit light using a single wide-angle coded pulse and captures the returning light from a range of locations simultaneously as pixels. If this light pulse enters a grape and experiences diffused scattering, each pixel of the ToF camera corresponding to the grape will receive some combination of light entering across the entire surface of the grape visible to the camera. In contrast, LiDARs typically build up the point cloud in a scanning process making measurements at a single scan point location at a time. This means that the light detected by the LiDAR may be more localised within the grape compared with the ToF camera. Given the different methods used by the two types of cameras, it is perhaps understandable then that each would have a different distortion pattern.

There have been a few reports of ToF cameras having a distance bias in fruit due to diffused scattering. Neupane et al. [37, 38] reported that ToF cameras provided distance measurements for mangoes, which were biased to be slightly too large, due to diffused scattering within the fruit. This distance bias increased over several days and was suggested as a means of measuring the ripeness of the mango fruit. Sarkar et al. [39] used this phenomenon to investigate the ripeness of apples using a ToF camera and polarisers. However, we have not seen any previous report of a shape distortion in ToF camera scans of fruit. The fact that the shape distortion is so pronounced for grapes may be due to the comparatively smaller size of the berries and relatively higher translucent properties compared to the other fruit that has been investigated previously.

This raises the question, could the distortion of RGB-D cameras that use ToF technology be used to provide a non-destructive estimation of grape properties such as ripeness? Future work is planned to investigate how the distortion effects vary with berry ripeness and size. This might also give some insight into the potential of correcting the ToF and LiDAR scans for these distortions in post-processing.

The ability to identify individual grapes from 3D scans could be beneficial. It potentially could allow the number and size of berries in bunches to be measured. Additionally, it might allow more accurate

yield estimation through 3D bunch architecture modelling. There have been several works that have used RANSAC to detect and size grapes. However, these works used high-resolution 3D scans captured using commercial laser and structured light scanners [5, 7, 8, 9, 10] and using photogrammetry [13], not depth cameras. Yin et al. [29] used RANSAC to fit cylinder shapes to the ZED RGB-D camera scans of grape bunches. However, this was related to the pose estimation of the entire grape bunch for robotic harvesting applications and did not attempt to fit individual grapes.

The RANSAC algorithm was used in this work on both the photogrammetry and RGB-D camera scans. The RANSAC algorithm showed some promise for detecting individual grapes in the RGB-D camera scans. All of the RGB-D cameras gave similar median 2D positions for the spheres/grapes relative to photogrammetry, as indicated in Table 4. However, the RANSAC algorithm produced fitted spheres with a smaller radius for the ToF and LiDAR cameras. This was to be expected given the shape distortion observed for these cameras.

The ability of RANSAC to correctly segment out individual berries was lower for the RGB-D cameras compared with that for the photogrammetry scans. As an example, in Figure 15, it can be seen that the Kinect V1 shows multiple grapes close to each other that have the same colour. This indicates that the algorithm has failed to separate these particular berries out as separate spheres. In contrast, a much higher proportion of the berries are correctly segmented for the photogrammetry scan.

The RANSAC algorithm also identified more grapes in the photogrammetry scans compared to that in the RGB-D camera. This is particularly pronounced for the grapes located around the edges of the bunch. However, this would appear to be mainly related to the way the photogrammetry scans are obtained using images captured from a range of positions relative to the grape bunch. The RGB-D camera images shown here in contrast are captured from a single location. This means the RGB-D cameras see a lower proportion of the surface area of the grape bunch. Improved results could be obtained by merging multiple RGB-D camera scans taken at a range of positions and angles relative to the grapes. This could be achieved using SLAM or a similar point cloud alignment technique [14]. This should then make the RGB-D camera scans more comparable to the photogrammetry scans.

Future Work

More investigation is needed to ascertain the optimal method of detection and sizing the grapes from RGB-D camera scans. Future work could look at fitting other shapes to the grape scans such as ellipsoids or a shape that is similar to the distortions due to diffused scattering effects for the ToF and LiDAR cameras. Additionally, custom-designed algorithms may be needed for these cameras. This may include correction of the distortion effects for these cameras.

The ToF and LiDAR cameras had slightly higher errors compared with the other two cameras indoors even when the grapes were painted or when the distance bias had been removed in post-processing. It is

possible that these errors could be reduced if additional filtering of the flying pixels was performed. However, this could potentially result in removing real scan points partially in the valleys between individual grapes. It is also possible that the error analysis process used here is overestimating the errors slightly for these cameras.

Improvements in the error analysis technique used in this work could also be performed. The error in the RGB-D cameras scans was obtained by comparing their depth scans with those obtained using photogrammetry. There could be some small errors in these photogrammetry scans. It appears that these scans had some smoothing in the valleys between grapes in a similar manner to the RealSense D415. It would be interesting in future work to use an alternative scanning system such as a commercial laser scanner for obtaining the ground truth scans.

The method used to calculate the distance errors could be improved in future work, particularly for the scans where a distance bias is present. One option could be to project a line from the location of the RGB-D camera to a scan point in its depth scan. One could then calculate the point on the line which is closest to a scan point on the photogrammetry scan (or where it passes through a mesh surface obtained from the photogrammetry scan). The distance along the line from that point to the RGB-D scan point could then be used as the depth error.

This work was performed with green grapes. Some preliminary testing with red grapes indicated that these also had a shape distortion and distance bias that appeared similar to that observed in the green grapes. However, this was not investigated in detail and more work is needed with other types of grapes.

The measurements described in this work were performed in controlled lab type environments. This was appropriate for the type of investigations performed in this study. However, it should be noted that achieving a fully automated system in a real vineyard environment would be more challenging. For example, this would require segmentation to allow automatic identification of grapes from leaves and stems [27]. There may also be occlusions by leaves or other grape bunches. More work is needed to address these types of challenges.

5. Conclusions

The Kinect V1 is no longer in production and hence is unlikely to be used in the future for grape yield estimation. However, it provides a comparison of the IR structure light technology with that used by other RGB-D cameras. The Kinect V1 was not able to function in direct sunlight. This is likely to be due to its projected IR pattern being saturated by sunlight. This indicates that RGB-D cameras that operate using IR structured light would only be suitable for measurements made at night or with a cover system that blocks out sunlight.

The Kinect V1 provided scans made indoors (out of direct sunlight) with relatively low errors for the

parts of the grapes facing the camera. However, it did not capture portions of the grapes, particularly in the valleys between individual grapes. While this might be adequate for rough volume estimations using a convex hull or mesh of the grape bunch scan, it does make identifying and sizing of individual grapes within the scan difficult. This is illustrated in the RANSAC results where the segmentation process struggled to correctly separate out many neighbouring grapes. In addition, it appears that the depth scans for the Kinect V1 had a relatively high quantisation compared with the other cameras.

The RealSense D415, which uses active stereoscopy, provided the lowest errors of the cameras analysed. Its indoor scans did not have the missing scan points or quantisation that was seen in the Kinect V1. However, it smoothed out the valleys between the grapes making it harder to detect individual grapes from the depth scans. The scans made with this camera in direct sunlight had slightly higher errors and missing scan points. In future work, we would look at adjusting the exposure of this camera in software to see if this issue can be addressed. However, it appears that sunlight was saturating its projected IR pattern, meaning it was acting purely as a passive stereo camera. This might indicate that cameras that operate using the AIRS technology may not have any additional benefit for yield estimation made in sunlight conditions compared with RGB-D cameras which operate using just passive stereo technologies. This may be investigated in future work.

The ToF (Kinect V2 and Kinect Azure) and LiDAR (Intel L515) cameras provided the best ability to detect individual grapes compared to the other cameras. However, they produced 3D scans of the grapes which were biased to give depth distances that were too large. Additionally, these cameras also produced distortions in the scans in the form of peaks centred on each grape location.

The distance bias and shape distortion were removed when the grapes were painted. This indicated that the distance bias and distortion were the results of diffused scattering within the grape. Previous work such as Neupane et al. [37] had reported measuring a distance bias for fruit using ToF cameras and have related this to the ripeness of the fruit. However, we are not aware of any previous studies which have reported a distortion in the shape of the scans of the fruit. It may be that this distortion is enhanced due to the small size of grape berries and their translucent properties.

The distance bias found in the LiDAR and ToF cameras scans of the grapes may not be an issue if one is only interested in the shape of the grape bunch. In fact, the distortion pattern makes it easier to identify individual grapes compared with the SL or AIRS cameras. However, more work is needed to investigate how much this distance bias and distortion affect the accuracy of grape volume/yield estimations. In our study, it did result in smaller detected berry diameters obtained using RANSAC compared with the other cameras. More work is needed to understand what factors such as ripeness, berry size, and variety play in the magnitude of the distance bias and shape of the distortion. With more understanding of these factors, it may be possible to use these distortions to perform non-destructive measurement of grape properties such as ripeness or possibly to correct for the distortions in post-processing.

In future work, we plan to investigate further the potential of the ToF and LiDAR cameras since they were less affected by sunlight and there is potential to utilise the distortion present in their scans for more accurately identifying individual berries. Additionally, there may be opportunities for using the distortion for non-destructive testing of berry properties.

References

[1] Laurent, C.; Oger, B.; Taylor, J.A.; Scholasch, T.; Metay, A.; Tisseyre, B. A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture. *Eur. J. Agron.* **2021**, *130*, 126339.

[2] Barriguinha, A.; de Castro Neto, M.; Gil, A. Vineyard yield estimation, prediction, and forecasting: A systematic literature review. *Agronomy* **2021**, *11*, 1789.

[3] Nuske, S.; Wilshusen, K.; Achar, S.; Yoder, L.; Narasimhan, S.; Singh, S. Automated visual yield estimation in vineyards. *J. Field Robot.* **2014**, *31*, 837–860.

[4] Zabawa, L.; Kicherer, A.; Klingbeil, L.; Töpfer, R.; Kuhlmann, H.; Roscher, R. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *164*, 73–83.

[5] Schöler, F.; Steinhage, V. Automated 3D reconstruction of grape cluster architecture from sensor data for efficient phenotyping. *Comput. Electron. Agric.* **2015**, *114*, 163–177.

[6] Tello, J.; Cubero, S.; Blasco, J.; Tardaguila, J.; Aleixos, N.; Ibanez, J. Application of 2D and 3D image technologies to characterise morphological attributes of grapevine clusters. *J. Sci. Food Agric.* **2016**, *96*, 4575–4583.

[7] Mack, J.; Schindler, F.; Rist, F.; Herzog, K.; Töpfer, R.; Steinhage, V. Semantic labeling and reconstruction of grape bunches from 3D range data using a new RGB-D feature descriptor. *Comput. Electron. Agric.* **2018**, *155*, 96–102.

[8] Rist, F.; Herzog, K.; Mack, J.; Richter, R.; Steinhage, V.; Töpfer, R. High-precision phenotyping of grape bunch architecture using fast 3D sensor and automation. *Sensors* **2018**, *18*, 763.

[9] Rist, F.; Gabriel, D.; Mack, J.; Steinhage, V.; Töpfer, R.; Herzog, K. Combination of an automated 3D field phenotyping workflow and predictive modelling for high-throughput and non-invasive phenotyping of grape bunches. *Remote Sens.* **2019**, *11*, 2953.

[10] Mack, J.; Rist, F.; Herzog, K.; Töpfer, R.; Steinhage, V. Constraint-based automated reconstruction of grape bunches from 3D range data for high-throughput phenotyping. *Biosyst. Eng.* **2020**, *197*, 285–305.

[11] Herrero-Huerta, M.; González-Aguilera, D.; Rodriguez-Gonzalvez, P.; Hernández-López, D. Vineyard yield estimation by automatic 3D bunch modelling in field conditions. *Comput. Electron. Agric.* **2015**, *110*, 17–26.

[12] Rose, J.; Kicherer, A.; Wieland, M.; Klingbeil, L.; Töpfer, R.; Kuhlmann, H. Towards automated large-scale 3D phenotyping of vineyards under field conditions. *Sensors* **2016**, *16*, 2136.

[13] Schneider, T.; Paulus, G.; Anders, K.H. Towards predicting vine yield: Conceptualization of 3D grape models and derivation of reliable physical and morphological parameters. *GI_Forum* **2020**, *8*, 73–88.

[14] Santos, T.; Bassoi, L.; Oldoni, H.; Martins, R. Automatic grape bunch detection in vineyards based on affordable 3D phenotyping using a consumer webcam. In Proceedings of the XI Congresso Brasileiro de Agroinformática (SBIAgro 2017), São Paulo, Brazil, 2–6 October 2017; pp. 89–98.

[15] Torres-Sánchez, J.; Mesas-Carrascosa, F.J.; Santesteban, L.G.; Jiménez-Brenes, F.M.; Oneka, O.; Villa-Llop, A.; Loidi, M.; López-Granados, F. Grape cluster detection using UAV photogrammetric point clouds as a low-cost tool for yield forecasting in vineyards. *Sensors* **2021**, *21*, 3083.

[16] Liu, S.; Whitty, M.; Cossell, S. A lightweight method for grape berry counting based on automated 3D bunch reconstruction from a single image. In Proceedings of the ICRA, IEEE International Conference on Robotics and Automation, Workshop on Robotics in Agriculture, Seattle, WA, USA, 25–30 May 2015; p. 4.

[17] Liu, S.; Zeng, X.; Whitty, M. 3DBunch: A novel iOS-smartphone application to evaluate the number of grape berries per bunch using image analysis techniques. *IEEE Access* **2020**, *8*, 114663–114674.

[18] Liu, S.; Zeng, X.; Whitty, M. A vision-based robust grape berry counting algorithm for fast calibration-free bunch weight estimation in the field. *Comput. Electron. Agric.* **2020**, *173*, 105360.

[19] Xin, B.; Liu, S.; Whitty, M. Three-dimensional reconstruction of *Vitis vinifera* (L.) cvs Pinot Noir and Merlot grape bunch frameworks using a restricted reconstruction grammar based on the stochastic L-system. *Aust. J. Grape Wine Res.* **2020**, *26*, 207–219.

[20] Fu, L.; Gao, F.; Wu, J.; Li, R.; Karkee, M.; Zhang, Q. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. *Comput. Electron. Agric.* **2020**, *177*, 105687.

[21] Marinello, F.; Pezzuolo, A.; Cillis, D.; Sartori, L. Kinect 3D reconstruction for quantification of grape bunches volume and mass. *Eng. Rural. Dev.* **2016**, *15*, 876–881.

[22] Hacking, C.; Poona, N.; Manzan, N.; Poblete-Echeverría, C. Investigating 2-D and 3-D proximal remote sensing techniques for vineyard yield estimation. *Sensors* **2019**, *19*, 3652–3671.

[23] Hacking, C.J. 2-D and 3-D Proximal Remote Sensing for Yield Estimation in a Shiraz Vineyard. Ph.D. Thesis, Stellenbosch University, Stellenbosch, South Africa, 2020.

[24] Kuan, Y.W.; Ee, N.O.; Wei, L.S. Comparative study of Intel R200, Kinect v2, and Primesense RGB-D sensors performance outdoors. *IEEE Sensors J.* **2019**, *19*, 8741–8750.

[25] Aquino, A.; Millan, B.; Diago, M.P.; Tardaguila, J. Automated early yield prediction in vineyards from on-the-go image acquisition. *Comput. Electron. Agric.* **2018**, *144*, 26–36.

[26] Kurtser, P.; Ringdahl, O.; Rotstein, N.; Berenstein, R.; Edan, Y. In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2031–2038.

[27] Kurtser, P.; Ringdahl, O.; Rotstein, N.; Andreasson, H. PointNet and geometric reasoning for detection of grape vines from single frame RGB-D data in outdoor conditions. In Proceedings of the 3rd Northern Lights Deep Learning Workshop (NLDL), Tromsö, Norway, 20–21 January 2019; Volume 1, pp. 1–6.

[28] Ivorra, E.; Sánchez, A.; Camarasa, J.; Diago, M.P.; Tardáguila, J. Assessment of grape cluster yield components based on 3D descriptors using stereo vision. *Food Control* **2015**, *50*, 273–282.

[29] Yin, W.; Wen, H.; Ning, Z.; Ye, J.; Dong, Z.; Luo, L. Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks. *Front. Robot. AI* **2021**, *8*, 626989.

[30] CloudCompare SOR (Statistical Outlier Removal) Filter. Available online: `https://www.cloudcompare.org/doc/wiki/index.php/SOR_filter` (accessed on 31 May 2022).

[31] Zinßer, T.; Schmidt, J.; Niemann, H. Point set registration with integrated scale estimation. In Proceedings of the International Conference On Pattern Recognition and Image Processing (PRIP 2005), Bath, UK, 22–25 August 2005; pp. 116–119.

[32] CloudCompare CCLib::ICPRegistrationTools Class Reference. Available online: `https://www.danielgm.net/cc/doc/CCLib/html/class_c_c_lib_1_1_i_c_p_registration_tools.html` (accessed on 31 May 2022).

[33] CloudCompare: Distances Computation. Available online: `https://www.cloudcompare.org/doc/wiki/index.php?title=Distances_Computation` (accessed on 31 May 2022).

[34] Schnabel, R.; Wahl, R.; Klein, R. Efficient RANSAC for point-cloud shape detection. *Comput. Graph. Forum* **2007**, *26*, 214–226.

[35] CloudCompare RANSAC Shape Detection (Plugin). Available online: `https://www.cloudcompare.org/doc/wiki/index.php/RANSAC_Shape_Detection_(plugin)` (accessed on 31 May 2022).

[36] Jennings, A. Matlab File Exchange: Sphere Fit (least Squared). Available online: `https://www.mathworks.com/matlabcentral/fileexchange/34129-sphere-fit-least-squared` (accessed on 31 May 2022).

[37] Neupane, C.; Koirala, A.; Wang, Z.; Walsh, K.B. Evaluation of depth cameras for use in fruit localization and sizing: Finding a successor to Kinect v2. *Agronomy* **2021**, *11*, 1780.

[38] Walsh, K.B.; Blasco, J.; Zude-Sasse, M.; Sun, X. Visible-NIR 'point' spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use. *Postharvest Biol. Technol.* **2020**, *168*, 111246.

[39] Sarkar, M.; Assaad, M.; Gupta, N. Phase based time resolved reflectance spectroscopy using time-of-flight camera for fruit quality monitoring. In Proceedings of the 2020 IEEE Sensors Applications Symposium (SAS), Kuala Lumpur, Malaysia, 9–11 March 2020; pp. 1–6.

Chapter 4

Grape Yield Estimation with a Smartphone's Colour and Depth Cameras using Machine Learning and Computer Vision Techniques

Grape Yield Estimation with a Smartphone's Colour and Depth Cameras using Machine Learning and Computer Vision Techniques

Baden Parr, Mathew Legg*, Fakhrul Alam

Department of Mechanical and Electrical Engineering, Massey University, Auckland, New Zealand

Abstract

A smartphone with both colour and time of flight depth cameras is used for automated grape yield estimation of Chardonnay grapes. A new technique is developed to automatically identify grape berries in the smartphone's depth maps. This utilises the distortion peaks in the depth map caused by diffused scattering of the light within each grape berry. This technique is then extended to allow unsupervised training of a YOLOv7 model for the detection of grape berries in the smartphone's colour images. A correlation coefficient (R^2) of 0.946 was achieved when comparing the count of grape berries observed in RGB images to those accurately identified by YOLO. Additionally, an average precision score of 0.970 was attained. Two techniques are then presented to automatically estimate the size of the grape berries and generate 3D models of grape bunches using both colour and depth information.

Keywords: grapes, yield estimation, berry detection, YOLO, depth camera, RGB-D

1. Introduction

Accurate grape yield estimation is crucial for wine growers since it enables them to effectively plan, organize, and take necessary actions, such as pruning and thinning, to optimize the quality of the wine they produce. Traditionally, yield estimation has been conducted through manual techniques such as visual observation or by cutting and weighing samples, which can be subjective, destructive, and time-consuming. Moreover, manual methods can result in undersampling of the vineyard, leading to potential errors. As a result, researchers are exploring automated yield estimation methods, mainly utilizing computer vision techniques [1, 2, 3].

Machine learning techniques have been used for detecting grape bunches in RGB (Red Green Blue) images. This has included convolutional neural networks [4] and different YOLO (You Only Look Once) models [5, 6, 7, 8]. However, for accurate yield volume estimations, it is desirable to count the number of berries within bunches and estimate the size of each berry.

*Corresponding author
Email addresses: 1badenparr@gmail.com (Baden Parr), M.Legg@massey.ac.nz (Mathew Legg), F.Alam@massey.ac.nz (Fakhrul Alam)

Several studies have detected individual grape berries in RGB images using spectral reflectance peaks in the images obtained using artificial lighting of the grapes in controlled field or lab conditions using smartphones [9, 10, 11] and camera systems [12, 13]. Machine learning has also been used to detect individual grape berries in RGB images. Coviello et al. [14] used dilated convolutional neural networks to count grapes in smartphone images. Miao et al. [15] used YOLOv3 to detect regions of interest around individual grapes. Additionally, a YOLO model for detecting individual grape berries can be downloaded from reference [16]. However, the training of these YOLO models will have been performed using manual labelling, which can be very time-consuming. Also, it would appear likely that this training would need to be repeated for different grape cultivar varieties.

For yield estimation, it is desirable to estimate the size of the individual berries within a bunch for accurate volume estimation. This is particularly the case for grape varieties that have a range of sizes within a bunch. Several studies have estimated the size of grapes and generated 3D models of grape bunches using Hough transforms to fit circles to grapes captured in camera and smartphone RGB images [17, 18, 19, 20]. These generally used backing boards to make the grapes more distinctive from the background and prevent circles from being detected in the background. Mirbod et al. [13] used spectral reflectance peaks to first detect the location of each grape berry and then used circle detection in this region to estimate the size of grapes in images. Miao et al. [15] also used a two-step process where a region of interest was identified around grape berries using a YOLOv3 model and then edge detection and ellipse fitting were used to estimate berry area in RGB images.

The size of the grape berries in an RGB image changes depending on the distance of the grapes from the camera due to perspective projection. One technique used to estimate the physical size of a grape berry from an RGB image is to place an object of known size next to the grapes. The size of an individual grape berry can then be obtained by comparing the size of the berry with the size of the reference object in the RGB image. Ang et al. [17, 18, 21] used a disk of known dimensions placed among the grapes or a checkerboard held next to the grapes to estimate the physical size of the grapes in a smartphone's or regular camera's images. Liu et al. [19, 20] also used a checkerboard image for this purpose. This allowed them to model the 3D structure of a grape bunch from 2D images. This process was extended by Xi et al [22, 23] to include constraint-based reconstructed grammars to "grow" the full 3D grape bunch structure from a single view 2D image.

The distance that a camera is from a checkerboard can be measured using the camera's intrinsic calibration parameters, which can be obtained using camera calibration software. This technique may have been used in the above works that used checkerboards.

It is desirable however not to have to use a reference object for estimating the size of grape berries for yield estimation. The physical size of grape berries can be estimated from their sizes in an RGB image if one knows the distance of the camera from the grapes when the image was taken, and one knows the camera's

calibration intrinsic parameters. Ivorra et al. [24] were able to estimate the size of grapes from RGB images without the need for a calibration object. They achieved this by measuring the distance that the camera was from the grapes using a stereo-depth camera. They used this distance combined with the size of grapes in the stereo camera's raw RGB images to estimate the physical size of grape berries. However, these results were obtained in controlled lab environments where the lighting, background, and camera position were carefully regulated, and manual refinement was required.

Grape size estimation and 3D modelling of grape bunches have also been performed using high-resolution 3D scans of grapes. This has included the use of photogrammetry. However, this involves a high computational load and can take significant time to process [25]. Stereo reconstruction has also been used to generate 3D models of grape bunches [26]. There has also been work using commercial high-resolution 3D scanners to generate 3D models of grapes in lab environments [27, 28]. However, these are expensive and do not seem suitable for practical use by farmers in the field.

There have been several studies that have used low-cost depth cameras to obtain 3D scans of grapes obtained using RGB-D (Red Green Blue - Depth) cameras for grape yield estimation. Marinello [29] and Hacking [30, 31] used the Microsoft Kinect V1 depth camera for yield estimation studies of grapes. This operates using infrared structured light, which did not work well in sunlight conditions due to the saturation of the projected infrared (IR) pattern. Kurtser et al. [32, 33] used an Intel RealSense D435 RGB-D camera for 3D scanning of grapes which uses active stereo. These works were not used to measure individual berry information. This is likely due to the relatively low resolutions of these cameras.

Parr et al. [34] compared the performance of several low-cost depth cameras for imaging grapes. It was shown that the ToF (Kinect V2 and Kinect Azure) and LiDAR (L515) depth cameras produced distortions in the 3D scans of individual grapes in the form of peaks centred on each grape location due to diffused scattering within the grapes. It was suggested that these distortions could be exploited to make the detection of grapes in ToF depth scans easier.

Previous research has employed smartphones to investigate grape yield estimation [35, 20, 19, 9, 18, 11]. An advantage of employing a smartphone in this context is that the majority of individuals already own one, thereby obviating the necessity for growers to invest in additional equipment. Many modern smartphones have built-in depth cameras in addition to RGB cameras. For example, the Samsung Galaxy Note 10+, Samsung Galaxy S20 Ultra, Huawei P30 Pro, etc. have built-in Time of Flight (ToF) cameras and the iPhone 12, 13 and 14 Pro and Pro Max models have built-in LiDARs. We are not aware of any previous works that have used the built-in depth cameras of a smartphone for grape yield estimation applications.

In this study, we utilize a Samsung Note 10+ smartphone to capture RGB images and ToF depth maps of Chardonnay grapes in field and lab environments. Grape detection is performed automatically by identifying distortion peaks in the ToF depth maps resulting from diffused light scattering within the grapes. We further train an unsupervised YOLOv7 model to detect the precise location of grape berries in RGB

images, leveraging the initial grape identification from the depth maps. Additionally, we develop techniques to estimate the size of grape berries and generate 3D models of grape bunches.

This article has the following contributions.

- We introduce a novel technique for the automatic detection of grape berry locations in 3D based on the peaks observed in the ToF depth maps captured by the smartphone. Building upon this technique, we extend it to enable unsupervised training of a YOLOv7 model for grape berry identification. To the best of our knowledge, this is the first instance of unsupervised training of a YOLO model specifically for grape detection, and we are not aware of any previous work that has employed a similar approach.
- The physical size of grape berries can be estimated from their size in the smartphone's RGB images using the distances from the camera to the grape berries that are automatically measured by the smartphone's depth camera. This removes the need for placing a calibration object next to the grapes, as has been done in previous work related to estimating berry size from RGB images captured in the field.
- A novel iterative modelling technique is introduced for estimating the sizes of grape berries based on their detected 3D positions, eliminating the need to estimate berry sizes from the RGB images. This approach offers an alternative method that does not rely on analyzing the RGB images to determine the berry sizes.

The remainder of the paper is organised as follows. Section 2 outlines the data collection methodology and processing used to generate RGB-D point clouds of grapes. The technique used to detect individual grapes from depth scans is described in Section 3. In Section 4, a technique used to train a YOLO model in an unsupervised manner is outlined. This model is then used to detect grape berries in the RGB images. The methods used to estimate the size of grape berries and perform 3D modelling of grape bunches are then presented in Section 5. Finally, the conclusion is presented in Section 6.

2. Methodology

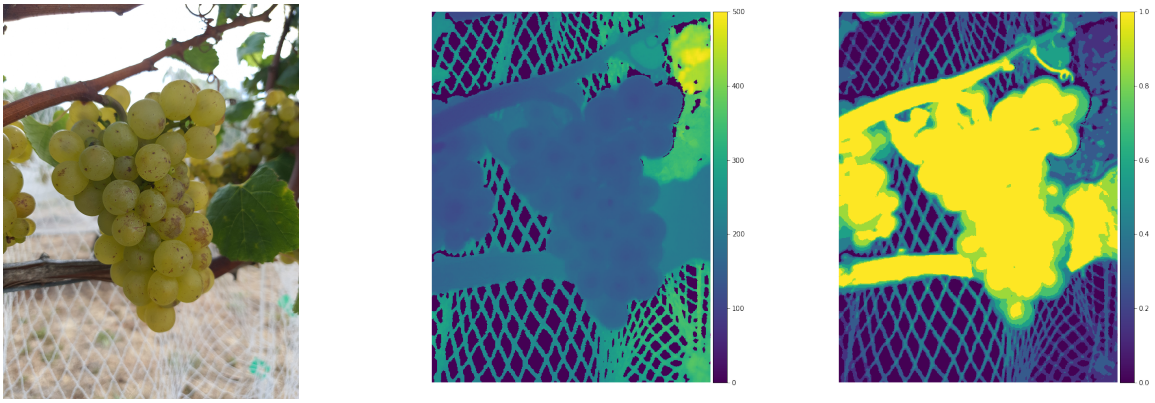
2.1. Data collection

Field measurements were made of Chardonnay grapes at the Villa Maria Estate in Auckland, New Zealand. These were performed about two weeks before harvest (late February). A Samsung Note 10+ smartphone was used to perform measurements on the grapes. This smartphone contains an RGB camera and a Time of Flight (ToF) depth camera. For each depth image, it also generates a confidence map, which provides an indication of the accuracy and validity of each point in the depth map.

At the time, there was no app available to capture depth map images from this camera. Therefore, a custom Android application was developed for this purpose. For each capture event, the application

automatically saved to file a 4032 × 3024 RGB image, a 640 × 480 depth map, and a time synchronised 640 × 480 confidence map. Additionally, a text file was also saved which contained the smartphone's GPS location and a reading from the smartphone's accelerometer taken at the time of capture.

Figure 1 provides an example of the RGB, depth, and depth confidence maps captured using this app for a grape bunch. (This grape bunch data will be used in most examples presented in this work for consistency.) The camera was able to capture depth maps in direct sunlight. No direct effort was made to take captures at any predetermined distance from the graph cluster. The only restriction was that each grape cluster should ideally fill the camera's frame. In total, 400 sets of images were captured of unique grape clusters throughout the vineyard.



(a) (b) (c)

Figure 1: Example images of the (a) RGB, (b) depth and (c) confidence maps captured by the Samsung Note 10+ of grapes in the field.

In order to build a YOLOv7 machine-learning model to identify grapes, a large number of scans of grapes were needed. To achieve this, 34 representative grape bunches were harvested from the vines and taken back to the lab. In turn, each grape bunch was suspended from a computer-controlled rotation table located 200 mm from the optical centre of the stationary Samsung Note 10+, see Figure 2. This distance was chosen to ensure that all grape bunches would fit within the camera's frame while being as close as possible. This methodology imitates our typical use of the phone's cameras when capturing images of bunches located on the vine.

The grape bunches were rotated through 360^o and the Samsung Note 10+ was used to capture an RGB image and depth and confidence maps at 10^o degree increments. Angles were not included where the structure of the rotation platform obscured the grape cluster. This resulted in a total of 1062 images of 34 grape bunches taken at a range of angles. Refer to Figure 3 for examples of scans captured using this technique. Additionally, 120 scans were taken from a range of angles in the lab of a potted grapevine, absent of grapes.

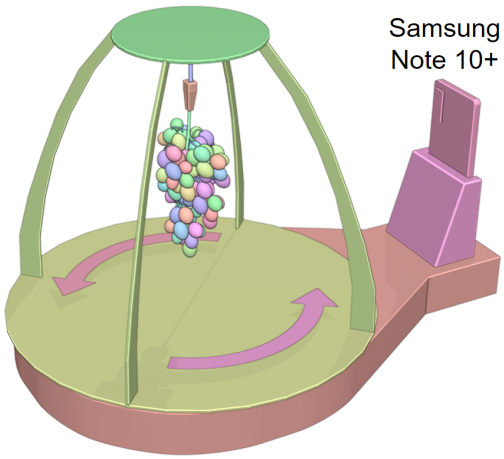
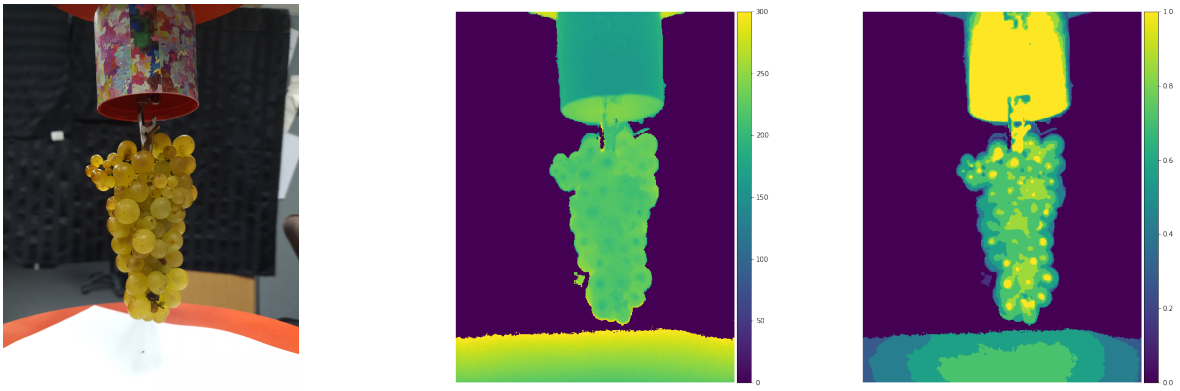


Figure 2: Diagram of the experimental setup where a turntable was used to capture images of a grape bunch using the smartphone from a range of angles.



(a) (b) (c)

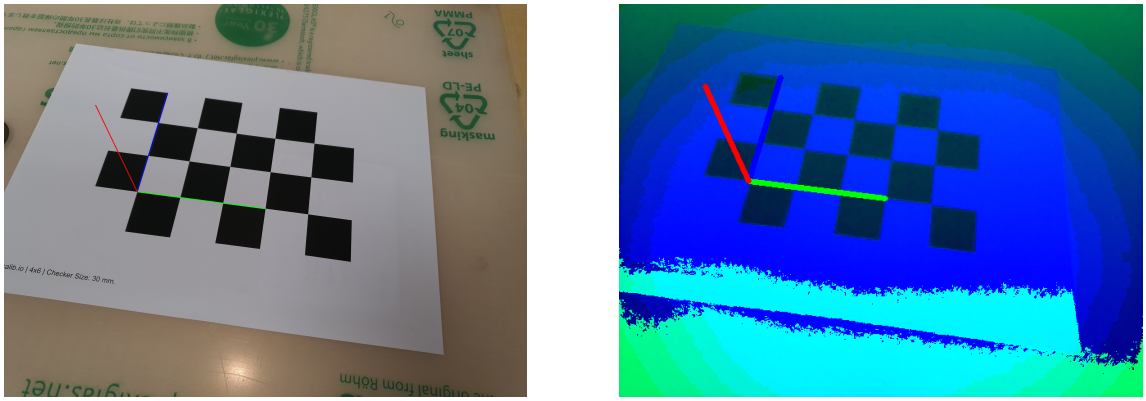
Figure 3: Examples of (a) an RGB image and (b) depth and (c) confidence maps captured by the Note 10+ of a grape bunch in the lab on a turntable. These were used to train a YOLO model to detect individual grapes.

2.2. Camera calibration

The Note 10+ cameras produced an RGB image and depth and confidence maps. In order to generate 3D-coloured depth point clouds (RGB-D) from these, the calibration parameters of the smartphone's cameras needed to be known. The smartphone's API did have calibration parameters stored. However, slight errors were found when using these to align the colour and depth maps when calculating the RGB-D point cloud. Therefore, a series of calibration colour and depth images were taken from a range of angles of a checkerboard pattern that was glued onto a sheet of acrylic. These measurements were made at similar distances from which the grape measurements were made. The black ink used to print the checkerboard pattern absorbed the infrared light emitted by the ToF camera meaning it showed up as voids (black) on the depth map.

This meant that the depth map images captured of the checkerboard could be used with camera calibration software.

The checkerboard images captured by the depth and colour cameras were separately calibrated using OpenCV v4.7.0. For this process, the RGB images were downsampled to be the same 640×480 resolution as the depth maps before calibration. This was done to reduce the processing burden and ease stereo registration. This 640×480 resolution will be used for RGB and depth images throughout the remainder of this work. Refer to Figure 4 for examples of corresponding depth and colour images obtained during this calibration with a common "real-world" reference frame shown. The estimated intrinsic parameters for both the colour and depth images were used along with the detected checkerboard coordinates for stereo calibration to obtain the extrinsic parameters defining the transformation from the RGB camera's reference frame to that of the depth camera.



(a) (b)

Figure 4: Examples of the (a) RGB and (b) depth calibration images captured by the smartphone's cameras.

2.2.1. Projecting between depth map and RGB images

Consider a pixel in the depth map with 2D coordinates $\bar{\boldsymbol{p}}_d$ in the X and Y axes directions, which has a depth value of Z. One can convert this into a 3D coordinate in the depth camera reference frame using

$$\bar{\boldsymbol{X}}_d = \begin{bmatrix} Z\ (\bar{\boldsymbol{p}}_d[1] - c_{d1})/f_{d1} \\ Z\ (\bar{\boldsymbol{p}}_d[2] - c_{d2})/f_{d2} \\ Z \end{bmatrix}, \tag{1}$$

where f_{d1} and f_{d2} are the depth camera's focal lengths in the X and Y axes directions, c_{d1} and c_{d1} are the coordinates of the central depth pixel in the depth map.

This 3D point $\bar{\boldsymbol{X}}_d$ can be moved from the depth camera's reference frame to the RGB camera's reference frame using the ridged body transformation

$$\bar{\boldsymbol{X}}_c = \boldsymbol{R}\ \bar{\boldsymbol{X}}_d + \bar{\boldsymbol{T}}, \tag{2}$$

where $\boldsymbol{R}$ is the stereo calibration rotation matrix and $\bar{\boldsymbol{T}}$ is the stereo translation vector.

This 3D point can be coloured by finding the colour of the corresponding pixel in the RGB image. The 3D point $\bar{\boldsymbol{X}}_c$ is first converted to normalised coordinates using

$$\bar{\boldsymbol{x_c}} = \begin{bmatrix} \bar{\boldsymbol{X}}_c[1]/Z \\ \bar{\boldsymbol{X}}_c[2]/Z \end{bmatrix}, \tag{3}$$

where $\bar{\boldsymbol{X}}_c[1]$ and $\bar{\boldsymbol{X}}_c[1]$ are respectively the X and Y axes components of $\bar{\boldsymbol{X}}_c$. This can be then converted into pixel coordinates on the RGB image using

$$\bar{\boldsymbol{p}} = \begin{bmatrix} f_{c1}\ \bar{\boldsymbol{x_c}}[1] + c_{c1} \\ f_{c2}\ \bar{\boldsymbol{x_c}}[2] + c_{c2} \end{bmatrix}, \tag{4}$$

where f_{c1} and f_{c2} are the RGB camera's focal lengths in the X and Y axes directions and c_{c1} and c_{c2} are the coordinates of the central pixel in the RGB image. No corrections were made for lens distortion or skew. The colour of this pixel can be used as the colour of the 3D point in either the RGB or depth camera's reference frames.

By repeating the above process for all pixels in the depth map, a coloured 3D point cloud can be generated. However, due to the perspective shift in some situations, multiple depth pixels will map to the same colour pixel. In this situation, only the point closest to the camera should be retained. Refer to Figure 5 for an example of the RGB image and the corresponding 3D coloured point cloud obtained using this method.

(a) (b)

Figure 5: Image (a) shows an example of the RGB photo captured by the smartphone's camera of a grape bunch in the field. Image (b) shows the corresponding colourised depth map.

3. Detection of berries in the ToF depth scans

Figure 6 shows an example of a ToF camera scan of a single grape before and after it has been sprayed with an opaque coating (AESUB 3D Scanning Spray). This illustrates how the diffused scattering of light within the grapes causes a distortion of the shape of the grape in the 3D scan. This manifests as a distinctive peak centred at the location of each grape. Observing this effect led to the idea that these peaks could potentially be used to facilitate the automatic detection of individual grape berries in ToF depth images [34].

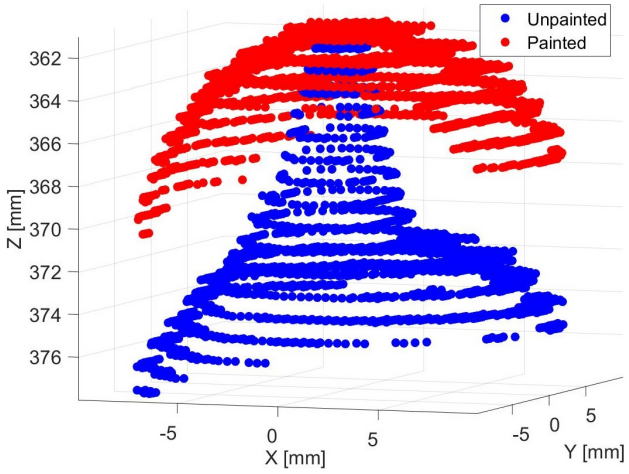


Figure 6: Example plot showing a peak in the depth map due to a grape that has been converted to 3D point cloud before and after it had been sprayed by an opaque coating. This illustrates how diffused scattering within the grape causes distortion of the depth scan in the form of peaks.

Figure 7 shows a block diagram of the technique used to investigate this idea. Each depth map captured by the smartphone ToF camera was filtered to reduce noise using the corresponding confidence map. Depth pixels that had a confidence value of less than 50% were removed. This had the primary effect of removing distant points. In all cases, the camera presented high confidence for pixels representing the grapes' surface. This 50% threshold was empirically determined from analysis of several images. Increasing this threshold caused the edges of grape clusters to erode slightly. Meanwhile, reducing the threshold caused background objects to be included and resulted in low persistence peaks to be detected due to the noise.

To identify potential grape locations, a peak detection algorithm was then used to identify peaks in the depth maps. A persistence homography technique [36, 37] was utilised for this due to its speed and robustness to noise. The persistence homography technique generated a persistence value for each identified peak, representing how significant a local maxima peak is in comparison to other local peaks.

Figure 8 (a) shows an example depth image of a grape cluster in the field, with the peaks detected by the persistence algorithm overlaid as white crosses. The algorithm is capable of detecting peaks that correspond to individual grapes. However, it also identifies peaks that correspond to the edges of grapes,

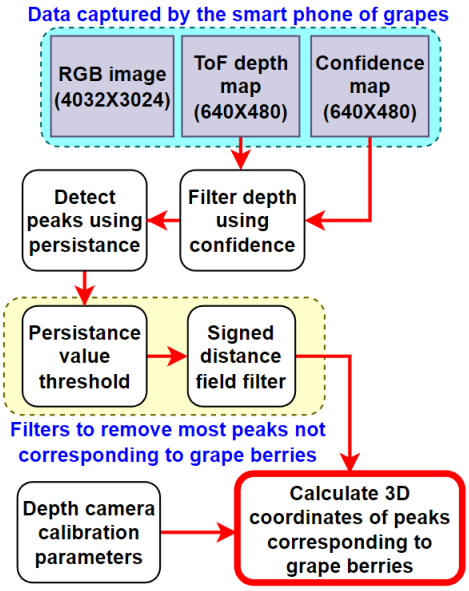
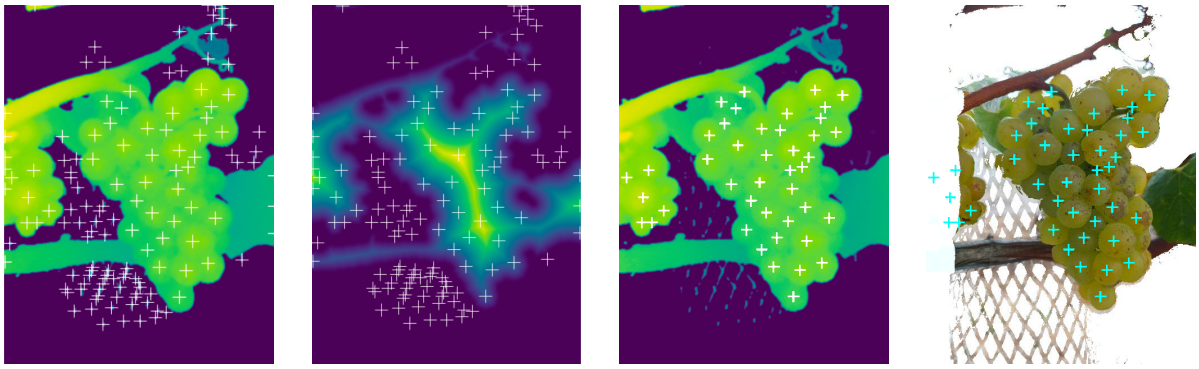


Figure 7: Block diagram showing the technique used for calculating the 3D coordinates of peaks in the depth map corresponding to grapes.

leaves, stems, and netting. Figure 8 (b) shows the Signed Distance Field (SDF) of this depth map, which was generated from a binary thresholded version of the depth map. This is utilized to remove peaks near edges by disregarding peaks that are closer than 7 pixels to an edge, as shown in Figure 8 (c) and (d). This threshold was chosen empirically to ensure only peaks close to the edge were removed and not those that may belong to small grapes. Future work will need to explore methods for scaling this threshold according to distance from the camera.

Manual analysis was performed for 50 of the scans of the grapes captured in the field. The total number of grapes visible in the images was manually counted. After which, peak locations were manually checked to see how many of the peaks corresponded to grapes and how many did not. Figure 9 (a) shows a plot of berries correctly detected (true positives) by the peak detection in depth maps relative to the total number of grapes visible in the corresponding RGB images. An R^2 value of 0.680 was obtained for the linear fit through this data. It can be seen that the technique underestimates the total number of grapes. Some grapes on the edge of the cluster were not detected, presumably because the centres of those grapes were occluded and therefore did not manifest as distinctive peaks in the depth map. In some cases, decreasing the SDF filtering threshold might result in an increase in the number of peaks being detected at the edges of the bunch. However, this will lead to an increase in the detection of peaks caused by other objects, such as leaves and netting, being erroneously identified as grapes.

The algorithm has been effective in eliminating most of the peaks that did not correspond to grapes. However, some incorrect peaks were detected, such as those corresponding to the peduncle between berries

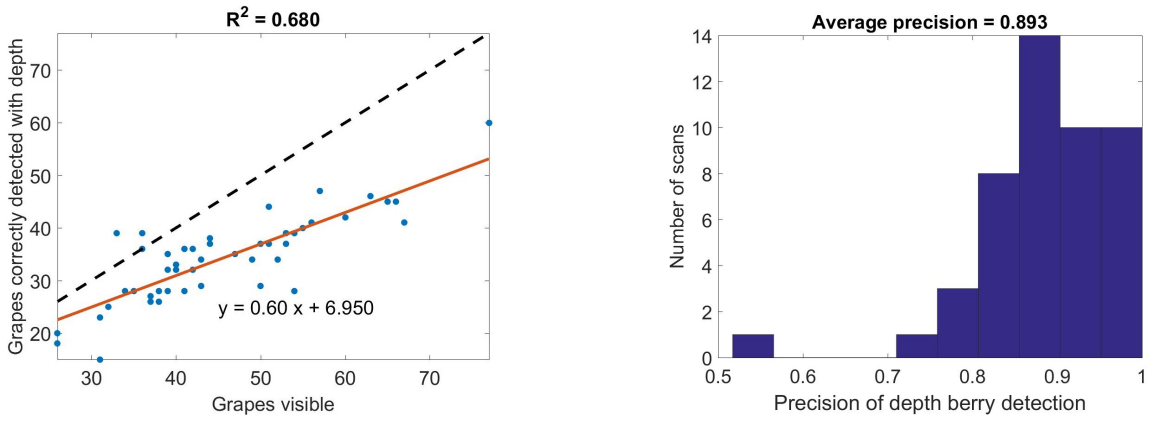


(a) (b) (c) (d)

Figure 8: These plots show the process of peak detection of a depth image captured in the field for the grape bunch shown in Figure 1. Plot (a) shows the peaks (white crosses) detected using persistence. Many peaks have been found on the netting in the background. In Plot (b), these peaks are shown over the generated signed distance field. Plot (c) shows the resulting peaks after signed distance field filtering was used with the aim of removing peaks not corresponding to grapes. Plot (d) shows these filtered peaks overlaid on the colourised depth map.

and on the rachis. Figure 9 (b) shows a histogram of the precision. The precision is calculated for each scan as the number of grapes correctly detected by the peak detection (true positives) divided by the total number of peaks identified as grapes (true positives plus false positives). An average precision of 0.893 was achieved.

The depth peak detection technique showed promise for automatically detecting grapes. However, it showed some limitations as described above. Work was therefore performed to investigate whether improved berry detection performance could be achieved by utilising the corresponding RGB images. This work is described in the following section.



(a) (b)

Figure 9: Plot (a) shows the relationship between the number of grapes correctly detected using peak detection in the depth maps relative to the total number of grapes counted manually in the corresponding RGB images. The identity line is shown as a dotted line and the line of best fit is shown in orange. Plot (b) shows a histogram of the precision.

4. Detection of individual berries in the RGB images

For this work, the popular YOLOv7 object detection model was chosen to facilitate the detection of grapes in smartphone's RGB images [38]. This selection was based on its well-established performance for object detection in complex images, as well as its pre-trained weights and open-source code that simplifies training new classes [39]. Training of a YOLO model requires images labelled in the form of bounding boxes around the object that the model is being trained to detect. This is traditionally done through supervised training; a process of manually selecting the bounds that encompass each instance of the object in question within an image. This process can be time-consuming, particularly for grape berry detection, which would require selecting individual grape berries in a large number of images, and would need to be repeated for distinct grape varieties [40]. Therefore, an automated technique was sought to perform unsupervised training utilising grapes detected through depth maps. The block diagram shown in Figure 10 illustrates the technique used to investigate this idea.

4.1. Dataset used for YOLO training

To reduce the potential of using false positives when automatically generating the bounding boxes, the scans of grapes captured in the lab were used for training, see Figure 3. Due to the controlled environment, the grape clusters could more easily be isolated from the image. The RGB images were downsampled to have the same 640×480 pixel resolution as the depth maps.

4.1.1. Bounding box generation using depth map data

To automatically generate bounding boxes in the RGB images used for YOLO training, the corresponding depth maps were employed. Firstly, the depth maps were filtered to isolate the grape clusters by removing

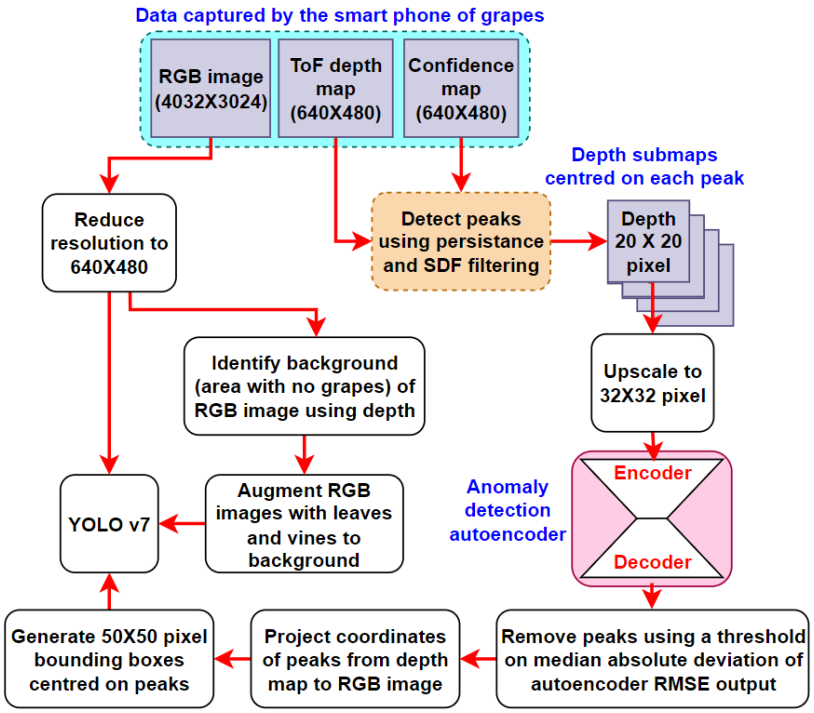


Figure 10: Block diagram of the technique used to perform unsupervised training of a YOLOv7 model for detection of individual berries in the RGB images using the estimated 3D coordinates from the depth maps.

points that were more than 300 mm away from the camera. This was chosen as the grapes were suspended 200 mm from the camera, and thus anything captured beyond 300 mm did not belong to the grape bunch. Next, the same technique explained in Section 3 was employed to detect peaks in the depth maps that corresponded to grapes. The confidence map with a threshold was applied to filter the depth map, following which the persistence algorithm was utilized to detect peaks. Finally, the signed distance field was used to eliminate peaks that were too close to the edges of the grape bunch.

4.1.2. Autoencoder based outlier rejection

As discussed in Section 3, the peak detection technique described would occasionally detect peaks that did not correspond to grapes. Inspection of these peaks showed they often related to the peduncle visible between berries or on the rachis where the clusters were hung. In each case, the erroneous peaks had significantly different profiles than the true positives, which themselves had relatively uniform shapes. See Figure 12 for examples of both cases. These false positives could influence the YOLO training and it was felt that a machine-learning technique could be used to detect these anomalies and filter out peaks that may not correspond to the centres of grape berries.

It was decided that an autoencoder would be used to identify peaks in the depth map that may not correspond to grapes. This decision was based on the idea that the autoencoder would be able to learn

information about the shape of different grape peaks, such as scaling factors and symmetries, making it effective for identifying outliers [41].

To ensure efficient training and minimize overfitting of potential outliers in the training set, the autoencoder's latent space was intentionally reduced in size. This reduction also aimed to prevent excessive complexity without generating artifacts in the reconstructed images.

The autoencoder was implemented using TensorFlow in Python. Figure 11 shows a block diagram of the autoencoder used in this work. The model consists of two parts: the encoder and the decoder. The encoder maps the input image to a lower-dimensional representation, while the decoder maps the encoded representation back to the original image.

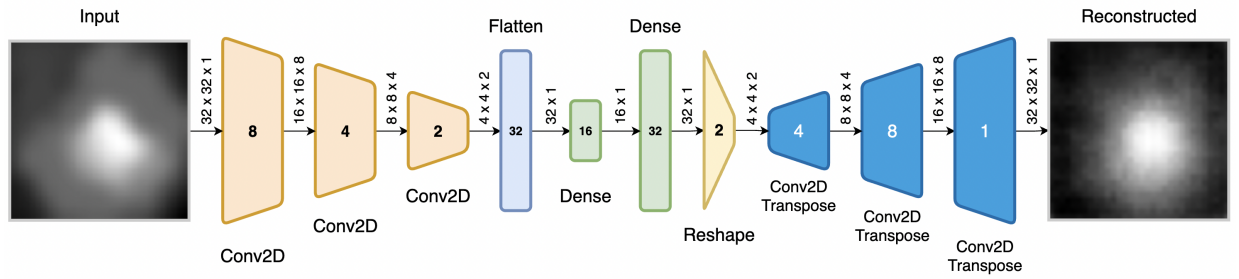
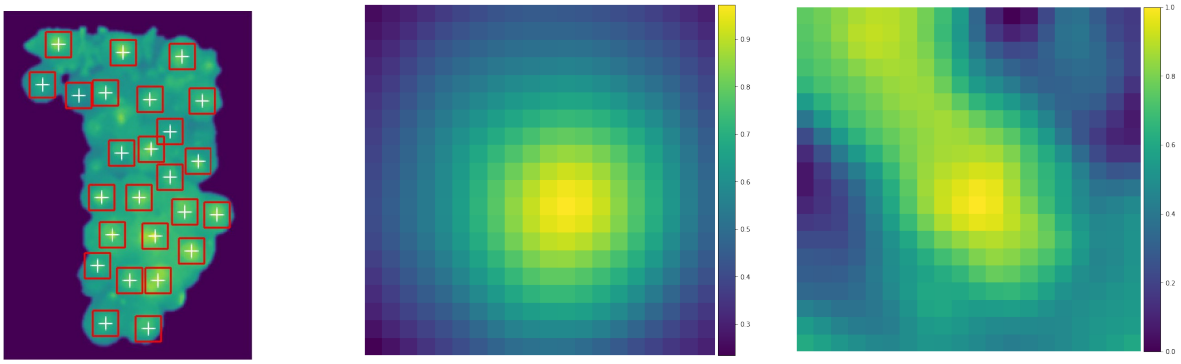


Figure 11: Block diagram of the autoencoder convolutional neural network.

To feed the autoencoder, a 20×20 pixel sub-map was taken from the depth map centred on the location of a detected peak, see Figure 12. This size was empirically chosen to be large enough to capture the majority of a peak's surface but not so large that it gets conflated by the surface of neighbouring grapes. In total, 33,844 of these sub-maps were generated and used to train the autoencoder. For convenience, each sub-map was then upsampled to a 32×32 resolution to make it a power of two suitable for use with the autoencoder.

The encoder takes the input image of size $32 \times 32 \times 1$ and applies three convolutional layers with 8, 4, and 2 filters respectively, each using a 4x4 kernel, stride of 2, and a ReLU activation function. This was designed to reduce the image size by half in subsequent layers, creating an effective encoding funnel for dimensionality reduction without relying on large dense layers.

The decoder takes the encoded representation as input and reconstructs the original image. The decoder starts with a fully connected layer with 32 units, followed by a reshape layer that transforms the output into a 4x4x2 tensor. Then, three transposed convolutional layers with 4, 8, and 1 filters, respectively, each using a 4x4 kernel, a stride of 2, and the ReLU activation function, are applied to the tensor. The last transposed convolutional layer has a sigmoid activation function, which maps the output to values between 0 and 1, representing the pixel intensities of the reconstructed image. The model takes the encoded representation as input and produces the reconstructed image as output. This model is trained using mean squared error



(a) (b) (c)

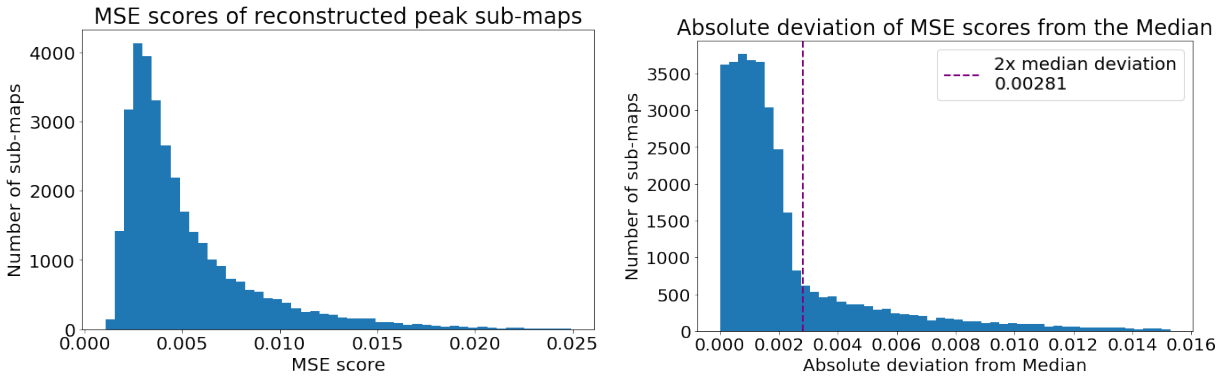
Figure 12: Plot (a) shows the 20 × 20 pixel sub-maps shown as red boxes surrounding the detected peaks in the depth map that are used as the inputs of the autoencoder. Plot (b) shows the average sub-map of the training set. Plot (c) shows an example of an erroneous sub-map relating to the peduncle visible in a cluster.

as the loss function between the original image and the reconstructed image.

Through empirical evaluation, the above architecture demonstrated optimal performance given the defined constraints and objectives. Decreasing the number of filters in each of the convolutional and transpose convolutional layers caused noticeable blocky artifacts in the reconstructions. Similarly, reductions in the latent space size (e.g., from 16x1 to 8x1) resulted in reconstructed images that exhibited similarity regardless of the input shape. These observations informed the decision to strike a balance between reducing complexity and preserving image fidelity. Future work will involve exploring alternative architectures to identify optimal designs.

After training, the MSE value generated by the autoencoder can be used with a threshold to classify if a peak corresponds to a grape or some other object (an anomaly). This threshold was determined by assessing the distribution of MSE scores of every sub-map in the dataset and filtering using the Median Absolute Deviation (MAD). The median of all scores was computed, and then the distance to this median was computed for all sub-maps. The threshold was set to two times the median of these distances, see 13. This allowed the autoencoder to be used as a strong filter to remove potential outlier peaks that might not correspond to the centre of the grapes.

The peaks remaining after the above filtering had been performed were then used to automatically generate bounding boxes in the RGB images for YOLO training. The coordinates of the peaks in the depth map were converted to coordinates in the corresponding RGB images using the stereo calibration parameters. Bounding box coordinates in the RGB image were then calculated using a 50 × 50 pixel square centred on the calculated peak location. This size was chosen to ensure that the grape was completely encompassed. Additionally, a second class label and bounding box were generated for the entire grape cluster based on the overall bounds of the detected grapes and an additional margin of 40 pixels.



(a) (b)

Figure 13: Plot (a) shows the distribution of mean square error (MSE) scores of all peaks that make up the training set for the autoencoder. Plot (b) shows the autoencoder's distribution of absolute deviations from the median along with the threshold used when applying these scores as a filter.

4.1.3. Background augmentation

A limitation of the lab-collected data set was that it did not include any images of leave or stems in them. This would have resulted in the YOLOv7 model not being applicable to the field trails. To address this, for each of the original turntable RGB images, two additional background-augmented images were added to the training data set. Each augmented RGB image was generated by taking an original turntable RGB image, isolating the grape bunch from the backgrounds using the depth map information, and overlaying the extracted grape bunch image over an RGB image captured of a grapevine randomly selected from a set of 120 images. The labels for the original source image were directly applied to these augmented images as the grape cluster itself remained unchanged. Examples of the two resulting images for one particular source image are shown in Figure 14.

4.2. YOLO Training

The dataset used to train and test the YOLOv7 model consisted of 3186 images labelled with grapes and grape clusters. The dataset was split into a training set (60%), a validation set (20%) and a test set (20%). The training process followed the method described in the official repository [39]. The default configuration parameters were used, and the training process was initiated with pre-trained weights provided in the official repository as "yolov7.pt". To keep memory requirements low, a batch size of 8 was used for training. The training process was run for a total of 20 epochs, and although more epochs were explored, no significant improvement was observed. The training process was completed in 0.615 hours using an Nvidia RTX 3090. Refer to Figure 15 for plots of the training results.

(a) (b)

Figure 14: Images of grapes captured on the turntable in the lab with images of leaves and vines added to the background.

4.3. YOLO results

The trained YOLO model was utilized to detect grapes in the RGB images captured in the field. Figure 16 presents examples of the detected grapes using the trained model. Most visible grape berries are accurately identified, although detection accuracy diminishes for out-of-focus grape bunches in the background. Additionally, some grape berries at the edge of the bunch remain undetected. Incorrect detection of netting, vines, or leaves as berries in the background also occur. Another issue arises when withered grapes are mistakenly identified as multiple grapes, as seen in 16 (c). This can be attributed to the absence of withered grapes during the YOLO model training process.

To evaluate the performance of YOLO for detecting individual berries, 50 field trial RGB images were randomly selected for manual analysis. (Note these RGB images corresponded to the same depth maps used for manual analysis of the depth peak detection technique shown in Figure 9.) These had a grape bunch centred in the image. Other grape bunches in the background were ignored in the analysis since generally either only a part of these secondary grape bunches could be seen or they were out of focus in the RGB images. Manual counting was then performed for the central grape bunch of the number of berries correctly and incorrectly detected by the YOLO model. These were then compared with the total number of grapes able to be manually counted in the grape bunch.

Figure 17 (a) compares the number of berries correctly detected using the YOLO model (true positives) to the total number of berries visible for each of the main grape bunches. There is a systematic underestimation of the number of berries counted using YOLO. Observations suggest that this is mainly due to missed grapes around the outside of the grape bunch, many of which have only a fraction of a berry visible. The fit through

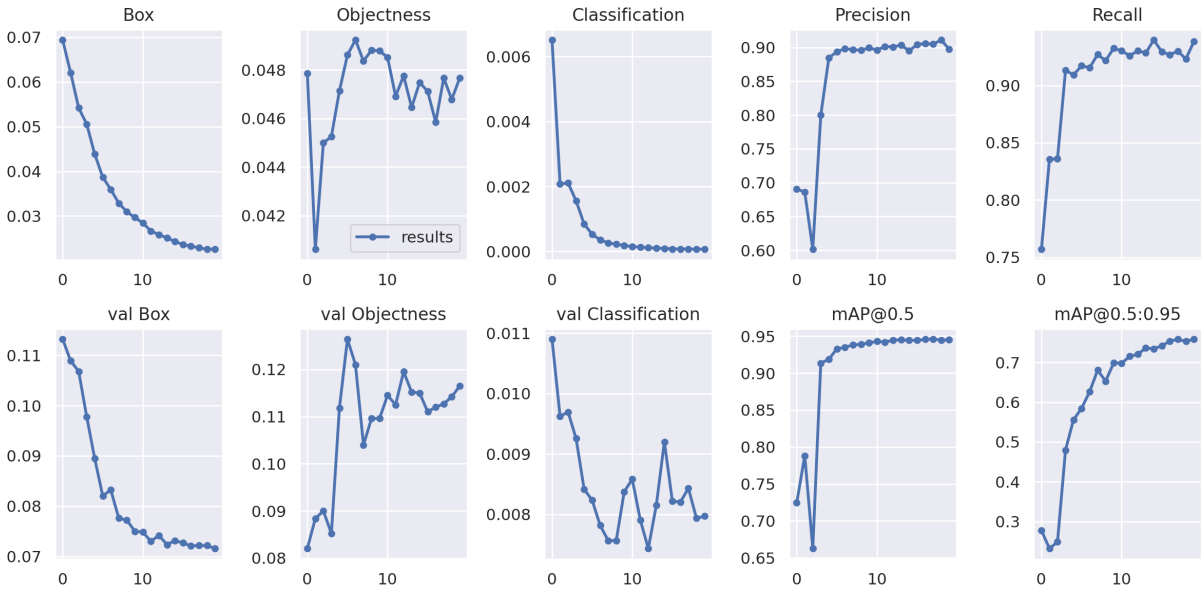


Figure 15: Results of YOLOv7 training over 20 epochs.

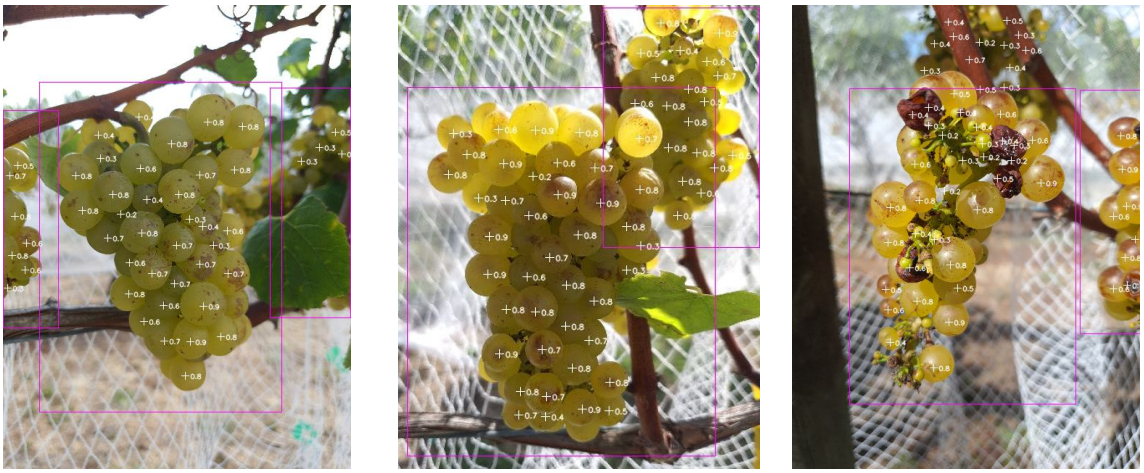
the data has a R^2 value of 0.946 and shows an increasing deviation from the one-to-one line as the number of berries in the cluster increase.

Figure 17 (b) shows a histogram of the precision. The precision for each scan is calculated from the number of berries counted by YOLO (true positives) divided by the sum of the total number of berries detected by YOLO (true positives + false positives). An average precision of 0.970 was achieved. The number of false positives within the bounded box selected by YOLO as the main grape bunch was 2.9% of the total number of visible grapes in the main bunch with only 12% of the images having more than 3 false positives.

4.4. Location of YOLO detections in 3D

The process of projecting the detected grape locations in an RGB image into 3D space is achieved by reversing the mapping process discussed in Section 2.2.1 to identify the closest corresponding depth pixel coordinate. However, due to differences in perspective and the way peaks align with the direction of measurement, these projected locations do not necessarily correspond to peaks in the point cloud. A seen in Figure 6 the peaks are the closes to the true surface of the grape. Therefore, using points from other areas on the surface will lead to significant errors in depth and subsequent estimated grape location.

To address this, a gradient descent technique was used to move the detected grape locations to the peaks in the depth map before projection. (Note that "gradient descent" is used rather than "gradient ascent" since the peaks were towards the camera and hence had lower depth values.) Specifically, the depth map scan of the grape bunch was filtered by removing pixels with corresponding confidence map values less than



(a) (b) (c)

Figure 16: Example photos from the field of grapes with YOLO detection of individual grapes berries overlaid as white crosses and confidence values. Also shown as magenta boxes is the YOLO detection of grape bunches. Plot (c) presents one of the more challenging images in the dataset where multiple withered grapes are visible and the trunk has been incorrectly labelled.

50% and smoothed using a 5×5 sliding average kernel. This reflects the confidence thresholding used earlier. An iterative gradient descent technique was then employed to move the grape locations to the top of the peaks. The 5x5 filter was selected empirically to provide suitable noise reduction ensuring the descent will not get stuck in small local minima but also retains definition so that individual peaks can be found. The effectiveness of this technique in improving the accuracy of grape location detection is illustrated in Figure 7.

The gradient descent algorithm adjusts the depth pixel location iteratively using the following formula:

$$\bar{\boldsymbol{p}}_{i+1} = \bar{\boldsymbol{p}}_i - \alpha \nabla \boldsymbol{J}(\bar{\boldsymbol{p}}_i) \tag{5}$$

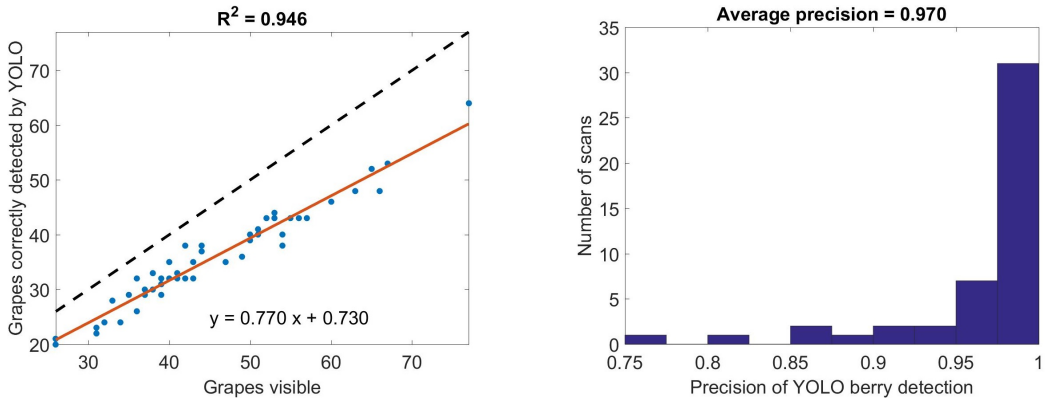
where $\bar{\boldsymbol{p}}_i$ is the pixel coordinate at the i^{th} iteration, $\alpha = 0.5$ is the traversal rate, and $\nabla \boldsymbol{J}(\boldsymbol{p}_i)$ is the gradient of the smoothed depth map $\boldsymbol{J}$ evaluated at coordinate $\bar{\boldsymbol{p}}_i$. This traversal rate was chosen empirically due to its stability and rate of convergence. Values significantly greater than this caused instabilities and values smaller caused convergence to take longer.

The gradient of the smoothed depth map with respect to the x and y axes is computed as follows:

$$\frac{\partial \boldsymbol{J}}{\partial x} = \frac{\boldsymbol{J}(y, x+1) - \boldsymbol{J}(y, x-1)}{2} \tag{6}$$

$$\frac{\partial \boldsymbol{J}}{\partial y} = \frac{\boldsymbol{J}(y+1, x) - \boldsymbol{J}(y-1, x)}{2} \tag{7}$$

where $\boldsymbol{J}(y, x)$ is the value of the smoothed depth map at pixel location (y, x). The pixel location is updated using the gradient descent formula until the algorithm terminates:



(a) (b)

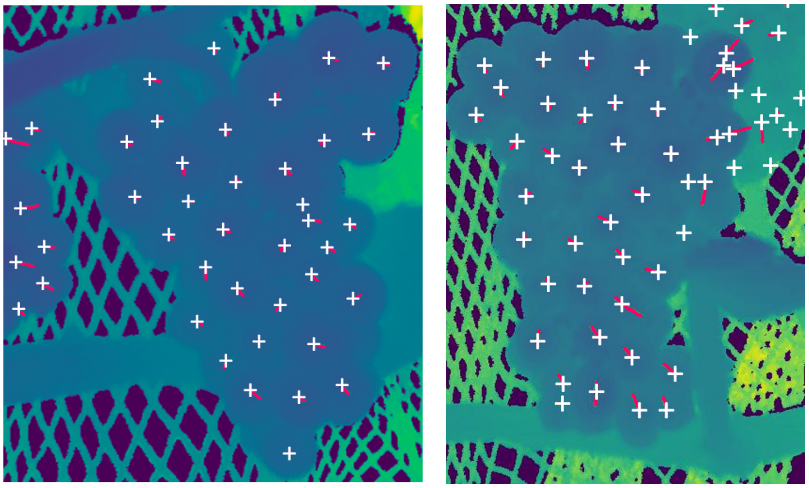
Figure 17: Plot (a) shows the number of grapes correctly detected by YOLO versus the number manually identified in the photos for 50 grape bunches. The identity line is shown as a dotted line and the line of best fit is shown in orange. Plot (b) shows a histogram of the precision.

$$\bar{\boldsymbol{p}}_{i+1} = \bar{\boldsymbol{p}}_i - \alpha \begin{bmatrix} \frac{\partial \boldsymbol{J}}{\partial y}(y,x) \\ \frac{\partial \boldsymbol{J}}{\partial x}(y,x) \end{bmatrix} \tag{8}$$

This was repeated for 50 iterations in order to move the berry location to the top of the nearest peak, see Figure 18. In all tested cases, this number of iterations was suitable to reach convergence. In cases where the traversal distance exceeded 15 pixels, the original coordinate was kept to prevent convergence on peaks too distant. This threshold was empirically determined to give the best results across the dataset.

In the majority of situations, this technique works well. However, in some edge cases, problems can show up. Some clear examples of these are demonstrated in Figure 18b. In some cases, the gradient descent process will cause multiple berry predictions to converge to the same peak within the depth map. This appears to happen most prominently on grapes that are occluded by a nearby grape causing the gradient to be stronger towards the peak of the occluding grape. In other cases, predictions of grapes behind the primary cluster (see the top right) ascend into the primary cluster. This is more common on grapes identified to the right and behind the primary cluster due to the parallax shift between the colour and depth sensors. This convergence behaviour also may help in some situations where the YOLO model predicts multiple grapes where only one exists. In this case, these predictions will converge to the same peak in the depth map. How these limitations can be solved or exploited will be the focus of future work.

Figure 19 compares the scans captured using the depth peak detection technique described in Section 3 and YOLO. We can see in the RGB image that there are slight differences between where the two methods have identified the location of the berries to be. However, for the 3D plot, gradient descent has been used to move the YOLO berry locations to the depth peaks. This results in similar berry locations being obtained using both methods for the 3D point cloud.



(a) (b)

Figure 18: Plots showing cropped versions of the depth maps corresponding to the grape bunches shown in Figure 16. The red lines show the path taken using the gradient descent technique to move from the berry locations obtained by YOLO to the peaks in the depth map, which are shown as white crosses.

(a) (b)

Figure 19: Cropped versions of the corresponding photo (a) and colourised depth map (b) of a grape bunch. Overlaid are the detected berry locations obtained using depth peak detection (red crosses) and the YOLO model (blue crosses). For the depth map, the YOLO berry locations were moved to local peaks using gradient descent.

5. Modelling of grape bunches

For grape yield estimation, it is desirable not only to count the number of grapes but also to be able to estimate the size of individual grapes so that grape volume can be estimated. This is particularly the case for grape varieties that typically have a wide range of berry diameters. The grapes used in this trial had a "hen and chicken" (Millerandage) effect where some grapes were smaller than others. Knowing the size distribution of grapes is a useful metric for effective vineyard management [42, 13]. Additionally, it is desirable to know the 3D structure of the grapes to allow better estimation of the total grape bunch volume and allow merging of scans of a grape bunch from multiple angles. Initial work was therefore conducted to estimate the size of the grapes detected and also construct a 3D model of the visible grapes.

5.1. Estimation of berry size from RGB images

The size of individual grape berries was detected from the RGB images using Hough transform circle detection. Initial trials using this technique over the entire RGB image gave poor results and were sensitive to hyper-parameter tuning; a limitation observed in existing works [17, 18, 21]. Therefore, a two-step process was adopted that exploits the available understanding of where berries are located. For each berry location detected by the YOLO model, a 480×480 pixel sub-image was extracted from the original high resolution 4032×3024 colour image, see Figure 20. This is similar to the technique that was used by Miao et al. [15].

This sub-image was converted to grayscale and edge detection was performed using a Sobel kernel. This kernel was then used to find the gradient at each pixel in the X and Y axes and the magnitude of these two obtained. Circles were then detected using a Hough transform. In cases where multiple distinct circles were detected, the circle closest to the berry location detected by the YOLO model was used. The process was repeated for all detected berry locations, see Figure 21.

The radius of the detected circles in pixels was able to be converted to a physical radius estimate using the knowledge of the distance of the camera from the grapes given by the depth camera and the camera calibration parameters. Similarly, the 3D location of each grape was also able to be estimated by projecting the YOLO detected locations onto the depth map using the process discussed in Section 4.4.

This information allowed a 3D model of the visible portion of the grape bunch to be generated. Spheres corresponding to the grapes were generated using their estimated size and 3D locations. This was done under the assumption that the peak found in the depth map corresponds to the closest point on the grape's surface to the camera. Furthermore, each grape can be modelled as a sphere where the point representing the peak is one of a pair of antipodal points, which, together with the camera origin and centre of the sphere, form a collinear set.

The 3D coordinate of the i_{th} sphere ($i = 1, \ldots, N$) is calculated using

$$\bar{\boldsymbol{C}}_i = \bar{\boldsymbol{X}}_i + \bar{\boldsymbol{d}}_i \; r_i \tag{9}$$

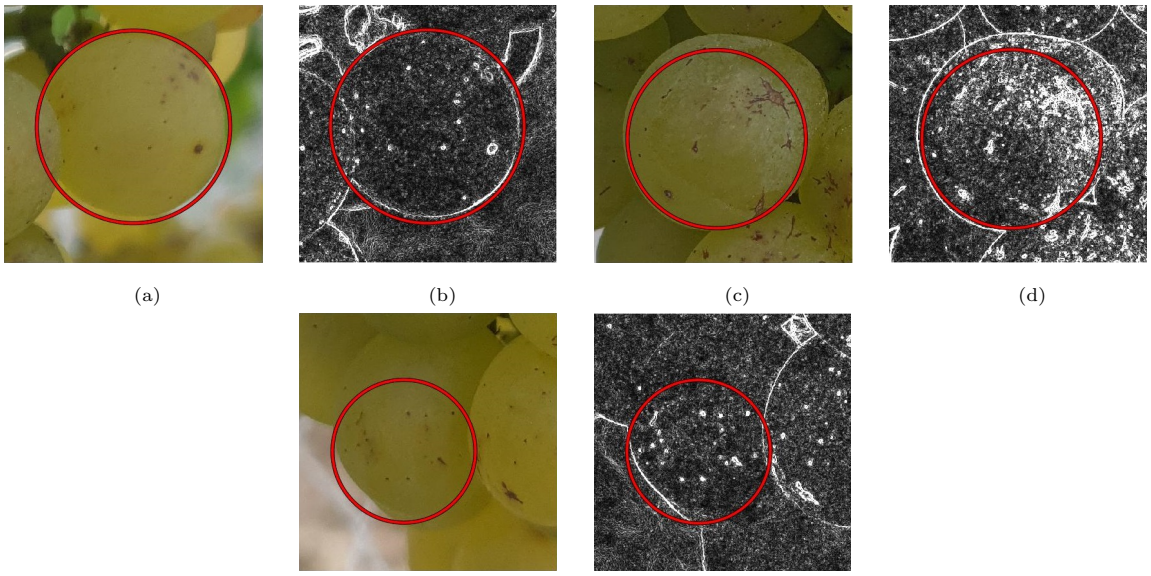


Figure 20: Photos (a), (c) and (e) show example RGB images that have been automatically cropped to be centred on a berry location identified by YOLO. Plots (b), (d) and (f) show the corresponding Sobel magnitude versions that emphasise edges. Overlaid are the detected circles obtained using a Hough transform on the Sobel filtered images. It can be seen that underestimation in the sizing of the grapes occurred due to factors such as the elliptical shape of the grapes and occlusion by neighboring grapes.

where $\bar{\boldsymbol{X}}_i$ is the 3D position of the detected peak, r_i is the radius of the sphere identified using circle detection, and $\bar{\boldsymbol{d}}_i$ is the normalized direction vector from the origin to the detected peak given by

$$\bar{\boldsymbol{d}}_{\boldsymbol{i}} = \frac{\bar{\boldsymbol{X}}_i}{\left\|\bar{\boldsymbol{X}}_i\right\|}. \tag{10}$$

Refer to Figure 22 for an example of a 3D model obtained using this technique overlain over the coloured 3D scan of the grapes generated from the depth and colour camera data.

This circle-fitting technique gives an approximation of the sizes of the grapes using the RGB images. However, errors can also be caused by the circles fitting to other features in the image such as the edge of another grape or colour changes on the surface of the grape. Also, many of the grapes appear as ellipses in the image rather than circles, which can lead to size estimation errors. Manual inspection of the fitted circles over the RGB images indicated that the circle fitting predominately resulted in some degree of underestimation compared to the true grape size. Refer to Figure 20 (c) and (e) for examples of this. Future work should explore using more advanced techniques such as the Holistically nested Edge Detection (HED) and ellipse fitting technique described in the work by Miao et al. [15].

These issues raised the question of if it is possible to estimate the size of the grapes without measuring their size from the image. The following section investigates this in more detail.

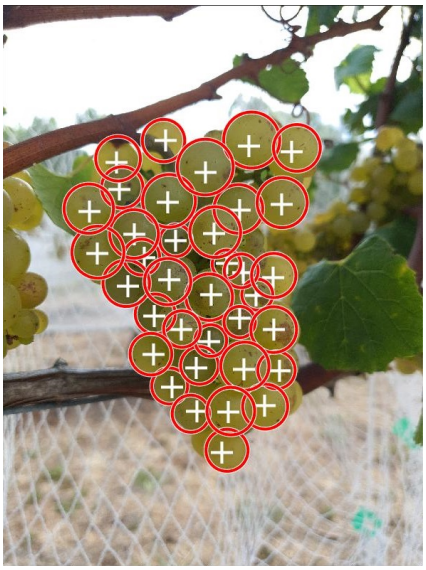


Figure 21: Sizing of grapes using circle detection.

5.2. Estimating berry size using depth

A technique was developed that estimates the size of the grapes and generates a 3D model using the identified locations of the grapes and the depth scan data rather than measuring the grape size from colour images. This approach works under the assumption that grape clusters are tightly packed and that they can be approximated as overlapping spheres. We also assume that the amount of overlap is proportional to their size.

Modelling of the 3D shape of the part of the grape bunch visible to the cameras was performed by creating a sphere for each grape using the method discussed in the previous section. To estimate the size of the grapes, the size of each sphere was iteratively adjusted with the aim of optimising the overlap between neighbouring spheres and limiting the maximum size to be within a limit realistic for grapes.

We want to optimize the maximum overlap distance between the sphere being optimised and the neighbouring spheres while keeping the maximum radius $r_{\max}$ of each sphere under a limit. For this work, this maximum radius was chosen to be 10 mm to ensure enough range to capture the largest berries we could expect in a bunch of Chardonnay grapes.

The maximum overlap of the i_{th} sphere with its k_{th} neighbour is determined by

$$\gamma_i = \max\{(r_i - r_k) - f(i,k)\} : \text{for } k = 1, \ldots, N, \tag{11}$$

where $f(i,k)$ is a function that returns the distance between the centres of the i_{th} and k_{th} spheres, and N is the total number of spheres. For each iteration, the algorithm calculates a change in radius Δr_i, for the i_{th} sphere based on the maximum overlap with neighbouring spheres. If the maximum overlap, γ_i is less than 50% of the sphere's current radius and the sphere's radius is less than $r_{\max}$, then the radius is increased

(a) (b)

Figure 22: Example of the 3D modelling of the grape bunch overlaid onto the colourised depth map shown in Figure 5. Plot (a) and (b) respectively show opaque and semi-transparent versions of the modelled spheres with diameters obtained using the circle fitting technique.

by a fixed amount of $\Delta r = 0.2$ mm. However, if the maximum overlap is larger than 50% of the sphere's current radius or the radius is over $r_{\max}$, then the radius is decreased by 10% of the current overlap. This can be expressed as

$$\Delta r_i = \begin{cases} 0.2, & \text{for } \gamma_i \leq 0.5 r_i \\ & \text{and } r_i \leq r_{\max} \\ -0.1\gamma_i, & \text{for } \gamma_i > 0.5 r_i \\ & \text{or } r_i > r_{\max} \end{cases}. \tag{12}$$

These thresholds and step sizes were chosen from empirical testing to help the simulation converge swiftly while also being stable. The 50% overlap threshold attempts to capture the squishing behaviour observed in the tightly grouped chardonnay bunches at the particular stage of development that images were captured. Different thresholds can be used to achieve different results and more work will need to be done to explore its impact on the simulations accuracy for different cultivars or stages of growth.

The simulation is run for a fixed number N_{iter} of iterations to ensure convergence. Changing the radius causes the position of the sphere to change, see Equation 9. Therefore, two passes over the spheres are conducted for each iteration. The first calculates Δr_i for each sphere, and the second applies this change and updates the centre of the sphere per Equation 9. The change in radius is applied to calculate the updated centre as follows

$$\bar{\boldsymbol{C}}_i[j+1] = \bar{\boldsymbol{d}}_i(\Delta r_i + r_i[j]) + \bar{\boldsymbol{X}}_i \tag{13}$$

where j ($j = 1, \ldots, N_{\text{iter}} - 1$) is the current iteration.

In this way, the position of the sphere is constrained by the relationship between its size and the amount of overlap with neighbouring spheres. The sphere will grow or shrink as necessary to avoid excessive overlap with neighbouring spheres, but it cannot exceed the maximum radius specified. The size of the spheres can then be used to estimate the size of the grape berries.

Refer to Figure 23 for an example of a model of a grape bunch using this technique. This method shows similar results compared to those obtained using the circle fitting technique shown in Figure 22. However, the resulting 3D scan does appear to be more accurate than the circle size approach when compared to the underlying colour image.

(a) (b)

Figure 23: Plots (a) and (b) respectively show opaque and semi-transparent versions of a 3D modelling of grapes generated by growing spheres at the location of grapes obtained from the peaks in the depth map. These are overlain over a 3D depth map scan of the grape bunch.

5.3. Comparison of grape sizes obtained using the RGB circle detection and depth techniques

Figure 24 presents a comparison of the grape sizes obtained using Hough transform circle fitting technique with those obtained using the depth technique for the grape bunch presented in Figures 22 and 23. It can be seen that the radii obtained using the RGB method was systematically lower than that obtained using the depth method. This is in line with expectations since the circle fitting tended to fit to one end of the ellipsoid shape of the grapes causing an systematic underestimation of the grape sizes, as illustrated in Figure 20 (c - f). Additionally, the distribution of these underestimations changes throughout the grape bunch depending on the shape or occlusion of individual berries. This explains some of the outliers present in the data and by extension the low correlation. In two cases, the simulated grape sizes have reached the maximum allowed

by the simulation, 10 mm. This indicates that those grapes were floating and did not have nearby grapes to constrain their size. In such a case, it may be more correct to use the RGB size estimations or a combination of the two. More work is needed to evaluate the performance of both of these methods against ground truth data and explore opportunities to combine both techniques for a robust solution.

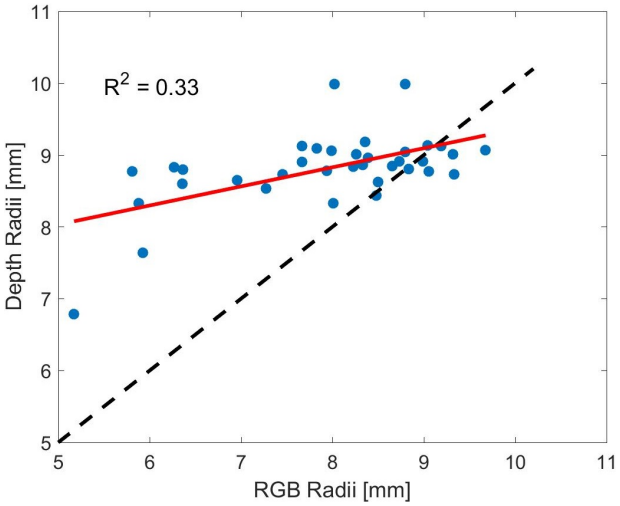


Figure 24: Comparison of the sizes of grapes obtained from circle detection in RGB images compared to those obtained using the peaks in the depth maps. The identity line is shown as a dotted line and the line of best fit is shown in red.

6. Conclusion

An Android app was developed for a Samsung Note 10+ smartphone to capture RGB images and depth and confidence maps simultaneously from its colour and ToF depth cameras. Stereo calibration of these two cameras was then performed using a checkerboard pattern. This allowed projection from the depth map to the corresponding RGB image along with mapping from the RGB image back to the depth map. Coloured 3D point clouds were able to be generated from the RGB and depth data. The colour in these point clouds was not utilised in this work but there is the potential for this to be used for improved results in future work.

The smartphone was used in field trials to perform scans of Chardonnay grapes in situ. Additionally, measurements were taken in the lab with samples of grape bunches from the field. A turntable was used to capture scans of each of these grape bunches at a range of angles.

A technique was developed to automatically identify grape berries in the depth maps using peak detection. This exploited the distortions in the ToF depth camera images due to diffused scattering within the berries. A persistence algorithm was used to detect peaks in the depth map. A signed distance field filter was used to remove peaks at the edges of objects and those corresponding to netting or leaves. This technique

successfully detected most of the visible grapes, though some were missing particularly at the edges. An R^2 value of 0.68 was obtained for a linear fit between the number of grapes visible in the RGB photos and those correctly detected using the depth peak fitting technique. An average precision of 0.893 was achieved.

Automatically identifying grape berries from peaks in the depth maps shows promise and further improvements could be made in future work. For example, the autoencoder that was developed for the YOLO training could be used to help improve the rejection of peaks that do not correspond to grapes. Including registered colour information in addition to depth could also help with improving the accuracy of this peak detection technique. Convolutional Neural Networks (CNN) may also provide an effective means of classifying which peaks are grapes.

A YOLOv7 model was trained to detect grape berries in RGB images captured by the smartphone. The dataset was constructed from lab-captured RGB and depth images. A technique was devised to facilitate unsupervised training by leveraging the peaks detected in the corresponding depth maps. An autoencoder was implemented to eliminate non-berry peaks, including those associated with visible rachis or peduncles. To enhance the dataset's adaptability to outdoor environments, training images were augmented with diverse foliage backgrounds through depth-based masking of grape clusters.

An R^2 value of 0.946 was achieved between a fitting of the number of berries correctly detected by YOLO and those manually counted in the RGB images and an average precision of 0.970 was achieved. The fit shows an underestimation in the number of berries detected by YOLO compared to those counted manually in the images. However, the strong relationship suggests that linear compensation would be an effective method of correction. The grapes that were missed by YOLO were mainly those around the edges of the grape bunch. This may be due to the fact that only a fraction of many of the berries on the edges of the bunch are clearly visible due to occlusions by other berries. However, it could also partly be related to the way the YOLO dataset was constructed and the low sensitivity of the peak detection process to occluded grapes on the edge of clusters. This may have meant that the YOLO model did not have sufficient training for grapes at the edge of the bunch. In future work, the manual selection of bounding boxes around berries missed by the peak detection could help improve the performance of the YOLO model in these cases. Additional training to remove YOLO detection of withered-up grapes could also be performed.

The YOLO model also struggled with grape bunches in the background where the RGB image was out of focus. In future work, this could be addressed using depth information by identifying the grape bunch of interest and filtering YOLO-detected points that would be out of focus in the RGB image. The grape bunch of interest could be identified based on its 3D position in the scan. Alternatively, one could manually click on the grape bunch in an image when capturing the scan using the app. One could also remove some of the false positives in the YOLO results using spatial filtering such as calculating the mean distance from the location of each detected berry to that of its K-nearest neighbours. More support could also be given to YOLO by producing augmented training data where grape bunches are blurred. Additionally, YOLO occasionally

produced false positive results by incorrectly identifying items in the background, such as netting or leaves, as grape berries. More varied background augmentations will help add robustness to these cases.

YOLO models are traditionally trained manually by labelling images by hand. However, this can be very time-consuming. For individual grape berry detection, this training would need to be repeated for different grape varieties. This is perhaps why only two works were found where YOLO has been used to detect individual grape berries [15, 16]. The automated approach introduced in this research, designed for unsupervised training of a YOLO model to detect grape berries, has the capacity to accelerate the training of YOLO models for a variety of grape types. The results presented here showed good accuracy. However, more work is needed to compare the accuracy obtained using this technique with that obtained using the traditional manual labelling method. Additionally, future work should investigate if adding manual labelling, particularly around the edges of grape bunches, could help improve the accuracy of the automated technique described in this work.

The berry locations detected by YOLO were able to be projected onto the depth map using the depth and RGB camera stereo calibration parameters. However, these predicted berry locations were generally slightly misaligned relative to the peaks in the depth map. A gradient descent technique was therefore developed that moved the projected YOLO berry locations to the top of nearby peaks. A potential issue with this approach is that it can result in two or more points detected by YOLO converging to the same 3D peak location. This can be seen demonstrated in Figure 18 (b) where berries detected in the background have ended up on a peak in the main bunch. Future work could investigate alternative methods of combining the presented YOLO and peak-based detection methods to provide a more robust approach to berry detection and filtering false positives.

Estimation of the size of grape berries in the grape bunches was performed with a two-step process. Firstly, circles were detected in the RGB images at the grape locations obtained using the YOLO model. Next, the physical size of each grape could be estimated from its size in the RGB image, the distance of the grape from the camera calculated from the depth map, and the RGB camera's intrinsic parameters. This eliminates the need for placing a reference object next to the grapes as has been used in previous works.

The generated size estimates were utilized to construct a 3D model of the grape bunch. By projecting the YOLO-detected berry locations from the RGB image onto the corresponding depth map, appropriately sized spheres were positioned at their respective 3D coordinates. Although this technique showed potential, it often underestimated berry sizes in our observations. In future work, one could investigate fitting ellipsoids to the data rather than spheres, as demonstrated in reference [15]. However, employing ellipses introduces additional hyper-parameters that significantly increase the transform space and may be influenced by image noise.

A sphere-growing optimisation technique was therefore developed to estimate the size of the berries in a grape bunch without having to measure their sizes in the RGB images. This approach works under the

assumption that grape clusters are tightly packed and that this can be approximated as overlapping spheres. Spheres were placed as before with the sizes iteratively adjusting to optimise the overlap among the entire cluster. This approach is sensitive to cases where grapes do not in reality touch other grapes or if some grapes are missed by the YOLO model. Future work could look at combining size estimates from RGB circle detection with this simulated approach as a method for constraining size expectations for each berry.

The results obtained using both the RGB and depths techniques showed promise. However, these results are qualitative. More work is needed in the future to compare these results with ground truth measurements of the physical sizes of the grape berries using callipers or scanning techniques such as laser scanners, photogrammetry etc.

Further work is also needed to build an understanding of the grapes within the cluster not visible to the camera. Past research has approximated these with a simple scaling factor. Our 3D models may also be accurate enough to extend to a complete phenotype estimate of the hidden structure, a process typically used with high-resolution 3D scans [27]. Additionally, scans from multiple angles may be able to be combined to increase the proportion of the grape bunch able to be included in the modelling.

This work was performed using a Samsung Note 10+ smartphone. However, the techniques should extend to any system that has a combined RGB camera and ToF or LiDAR depth cameras. This includes a range of modern smartphones from Apple and Android and low-cost depth camera systems that are currently commercially available such as the Microsoft Azure Kinect DK. Additionally, the YOLO model is suitable for berry detection with standalone RGB cameras without a smartphone for depth sensors.

The data used in this work was for Chardonnay grapes which are green in colour. Initial lab-based trials were also performed on red table grapes, though the results are not presented here. Similar peaks were observed in the ToF depth maps captured of these red grapes compared to those presented in this work. This leads to some confidence that the technique would be extendable to other grape varieties. However, additional experiments with different grape varieties would be beneficial.

The field trials were performed on the grapes approximately two weeks before harvest. It is beneficial for growers to perform yield estimation measurements at this stage of growth so that they can estimate the volume of grapes that will be harvested, etc. However, it is also desirable to be able to perform yield estimations at different stages of grape maturity. It is likely that the diffused scattering within the grapes that is causing the peak distortion may change with grape maturity. Therefore, it would be desirable in future work to perform further trials of grapes at a range of maturity levels.

7. Acknowledgement

The researchers would like to acknowledge Bragato Research Institute (a subsidiary of New Zealand Winegrowers) as this research was supported in part by the Rod Bonfiglioli Scholarship.

References

[1] A. Barriguinha, M. de Castro Neto, A. Gil, Vineyard yield estimation, prediction, and forecasting: A systematic literature review, Agronomy 11 (9) (2021) 1789.

[2] C. Laurent, B. Oger, J. A. Taylor, T. Scholasch, A. Metay, B. Tisseyre, A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture, European Journal of Agronomy 130 (2021) 126339.

[3] H. Moreno, D. Andújar, Proximal sensing for geometric characterization of vines: A review of the latest advances, Computers and Electronics in Agriculture 210 (2023) 107901. doi:https://doi.org/10.1016/j.compag.2023.107901.

[4] T. T. Santos, L. L. de Souza, A. A. dos Santos, S. Avila, Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association, Computers and Electronics in Agriculture 170 (2020) 105247.

[5] H. Li, C. Li, G. Li, L. Chen, A real-time table grape detection method based on improved yolov4-tiny network in complex background, Biosystems Engineering 212 (2021) 347–359.

[6] R. Zhao, Y. Zhu, Y. Li, An end-to-end lightweight model for grape and picking point simultaneous detection, Biosystems Engineering 223 (2022) 174–188.

[7] B. Liu, L. Luo, J. Wang, Q. Lu, H. Wei, Y. Zhang, W. Zhu, An improved lightweight network based on deep learning for grape recognition in unstructured environments, Information Processing in Agriculture (2023).

[8] L. Shen, J. Su, R. He, L. Song, R. Huang, Y. Fang, Y. Song, B. Su, Real-time tracking and counting of grape clusters in the field based on channel pruning with yolov5s, Computers and Electronics in Agriculture 206 (2023) 107662. doi:https://doi.org/10.1016/j.compag.2023.107662.

[9] M. Grossêtete, Y. Berthoumieu, J.-P. Da Costa, C. Germain, O. Lavialle, G. Grenier, A new approach on early estimation of vineyard yield: Site specific counting of berries by using a smartphone, in: European Conference on Precision Agriculture, 2011, pp. 8–pages.

[10] M. Grossetete, Y. Berthoumieu, J.-P. Da Costa, C. Germain, O. Lavialle, G. Grenier, et al., Early estimation of vineyard yield: Site specific counting of berries by using a smartphone, in: International Conference of Agricultural Engineering—CIGR-AgEng, 2012.

[11] A. Aquino, I. Barrio, M.-P. Diago, B. Millan, J. Tardaguila, vitisBerry: An Android-smartphone application to early evaluate the number of grapevine berries by means of image analysis, Computers and Electronics in Agriculture 148 (2018) 19–28.

[12] D. Font, T. Pallejà, M. Tresanchez, M. Teixidó, D. Martinez, J. Moreno, J. Palacín, Counting red grapes in vineyards by detecting specular spherical reflection peaks in rgb images obtained at night with artificial illumination, Computers and Electronics in Agriculture 108 (2014) 105–111.

[13] O. Mirbod, L. Yoder, S. Nuske, Automated measurement of berry size in images, IFAC-PapersOnLine 49 (16) (2016) 79–84.

[14] L. Coviello, M. Cristoforetti, G. Jurman, C. Furlanello, Gbcnet: In-field grape berries counting for yield estimation by dilated cnns, Applied Sciences 10 (14) (2020) 4870.

[15] Y. Miao, L. Huang, S. Zhang, A two-step phenotypic parameter measurement strategy for overlapped grapes under different light conditions, Sensors 21 (13) (2021) 4532.

[16] new-workspace hzmvk, Berry_yoloV5 Dataset, https://universe.roboflow.com/new-workspace-hzmvk/berry_yolov5-slwnw, last visited on 20 March 2023 (nov 2021).

[17] L.-M. Ang, K. Seng, A. Oczkowski, A. Deloire, L. Schmidtke, Development of a smartphone app for berry quality assessment, in: Vigne et Vin Publications, 2018, pp. 79–85.

[18] L. Schmidtke, Developing a phone-based imaging tool to inform on fruit volume and potential optimal harvest time, Tech. Rep. Project No. CSU 1501, National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga, New South Wales, Australia (June 2018).

[19] S. Liu, X. Zeng, M. Whitty, A vision-based robust grape berry counting algorithm for fast calibration-free bunch weight estimation in the field, Computers and Electronics in Agriculture 173 (2020) 105360.

[20] S. Liu, X. Zeng, M. Whitty, 3dbunch: A novel ios-smartphone application to evaluate the number of grape berries per bunch using image analysis techniques, IEEE Access 8 (2020) 114663–114674.

[21] National Wine and Grape Industry Centre, WineOz SmartGrape, https://play.google.com/store/apps/details?id=com.nwgic.grapeyield, (Last visited on 5 Sept 2022.) (2019).

[22] B. Xin, S. Liu, M. Whitty, Three-dimensional reconstruction of vitis vinifera (l.) cvs pinot noir and merlot grape bunch frameworks using a restricted reconstruction grammar based on the stochastic l-system, Australian Journal of Grape and Wine Research 26 (3) (2020) 207–219.

[23] B. Xin, M. Whitty, A 3D grape bunch reconstruction pipeline based on constraint-based optimisation and restricted reconstruction grammar, Computers and Electronics in Agriculture 196 (2022) 106840. doi:https://doi.org/10.1016/j.compag.2022.106840.

[24] E. Ivorra, A. Sánchez, J. Camarasa, M. P. Diago, J. Tardáguila, Assessment of grape cluster yield components based on 3D descriptors using stereo vision, Food control 50 (2015) 273–282.

[25] J. C. Rose, A. Kicherer, M. Wieland, L. Klingbeil, R. Töpfer, H. Kuhlmann, Towards automated large-scale 3D phenotyping of vineyards under field conditions, Sensors 16 (12) (2016). doi:10.3390/s16122136.

[26] M. Herrero-Huerta, D. González-Aguilera, P. Rodriguez-Gonzalvez, D. Hernández-López, Vineyard yield estimation by automatic 3d bunch modelling in field conditions, Computers and Electronics in Agriculture 110 (2015) 17–26. doi:https://doi.org/10.1016/j.compag.2014.10.003.

[27] F. Schöler, V. Steinhage, Automated 3d reconstruction of grape cluster architecture from sensor data for efficient phenotyping, Computers and Electronics in Agriculture 114 (2015) 163–177.

[28] J. Mack, F. Schindler, F. Rist, K. Herzog, R. Töpfer, V. Steinhage, Semantic labeling and reconstruction of grape bunches from 3d range data using a new rgb-d feature descriptor, Computers and Electronics in Agriculture 155 (2018) 96–102. doi:https://doi.org/10.1016/j.compag.2018.10.011.

[29] F. Marinello, A. Pezzuolo, D. Cillis, L. Sartori, et al., Kinect 3D reconstruction for quantification of grape bunches volume and mass, Engineering for Rural Development 15 (2016) 876–881.

[30] C. Hacking, N. Poona, C. Poblete-Echeverría, Vineyard yield estimation using 2-D proximal sensing: a multitemporal approach, OENO One 54 (4) (2020) 793–812.

[31] C. J. Hacking, 2-D and 3-D proximal remote sensing for yield estimation in a Shiraz vineyard, Ph.D. thesis, Stellenbosch: Stellenbosch University (2020).

[32] P. Kurtser, O. Ringdahl, N. Rotstein, H. Andreasson, Pointnet and geometric reasoning for detection of grape vines from single frame RGB-D data in outdoor conditions, in: 3rd Northern Lights Deep Learning Workshop, Tromsö, Norway, Vol. 1, NLDL, 2020, pp. 1–6.

[33] P. Kurtser, O. Ringdahl, N. Rotstein, R. Berenstein, Y. Edan, In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera, IEEE Robotics and Automation Letters 5 (2) (2020) 2031–2038.

[34] B. Parr, M. Legg, F. Alam, Analysis of depth cameras for proximal sensing of grapes, Sensors 22 (11), dio:10.3390/s22114179 (2022).
URL https://www.mdpi.com/1424-8220/22/11/4179

[35] J. Tardaguila, M. Stoll, S. Gutiérrez, T. Proffitt, M. P. Diago, Smart applications and digital technologies in viticulture: A review, Smart Agricultural Technology 1 (2021) 100005.

[36] S. Huber, Persistent homology in data science, in: P. Haber, T. Lampoltshammer, M. Mayr, K. Plankensteiner (Eds.), Data Science – Analytics and Applications, Springer Fachmedien Wiesbaden, Wiesbaden, 2021, pp. 81–88.

[37] S. Huber, Topological peak detection in two-dimensional data, available online: https://www.sthu.org/code/

codesnippets/imagepers.html (2022).

[38] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, arXiv preprint arXiv:2207.02696 (2022).

[39] C.-Y. Wang, Official yolov7, https://github.com/WongKinYiu/yolov7, (Last visited on 18 May 2023.) (2022).

[40] T. A. Ciarfuglia, I. M. Motoi, L. Saraceni, M. Fawakherji, A. Sanfeliu, D. Nardi, Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data, Computers and Electronics in Agriculture 205 (2023) 107624. doi:https://doi.org/10.1016/j.compag.2023.107624.

[41] Z. Zhu, X. Wang, S. Bai, C. Yao, X. Bai, Deep learning representation using autoencoder for 3D shape retrieval, Neurocomputing 204 (2016) 41–50, big Learning in Social Media Analytics. doi:https://doi.org/10.1016/j.neucom.2015.08.127. URL https://www.sciencedirect.com/science/article/pii/S0925231216301047

[42] J. M. Miras-Ávalos, I. Buesa, A. Yeves, D. Pérez, D. Risco, J. R. Castel, D. S. Intrigliolo, et al., Unravelling the effects of berry size on'tempranillo'grapes under different field practices, Ciência e Técnica Vitivinícola 34 (1) (2019) 1–14.

Chapter 5

Occluded Grape Cluster Detection and Vine Canopy Visualisation Using an Ultrasonic Phased Array

This chapter is republished in accordance with MDPI's copyright policy. The work presented here is the accepted version of the published article. Therefore, the contents are the same but there may be stylistic differences to the published article.

 B. Parr, M. Legg, S. Bradley, F. Alam. Occluded grape cluster detection and vine canopy visualisation using an ultrasonic phased array. Sensors (2021) 21(6), 2182. doi.org/10.3390/s21062182

Occluded Grape Cluster Detection and Vine Canopy Visualisation Using an Ultrasonic Phased Array

Baden Parr, Mathew Legg*, Fakhrul Alam

Department of Mechanical and Electrical Engineering, Massey University, Auckland, New Zealand

Abstract

Grape yield estimation has traditionally been performed using manual techniques. However, these tend to be labour intensive and can be inaccurate. Computer vision techniques have therefore been developed for automated grape yield estimation. However, errors occur when grapes are occluded by leaves, other bunches, etc. Synthetic aperture radar has been investigated to allow imaging through leaves to detect occluded grapes. However, such equipment can be expensive. This paper investigates the potential for using ultrasound to image through leaves and identify occluded grapes. A highly directional low frequency ultrasonic array composed of ultrasonic air-coupled transducers and microphones is used to image grapes through leaves. A fan is used to help differentiate between ultrasonic reflections from grapes and leaves. Improved resolution and detail are achieved with chirp excitation waveforms and near-field focusing of the array. The overestimation in grape volume estimation using ultrasound reduced from 222% to 112% compared to the 3D scan obtained using photogrammetry or from 56% to 2.5% compared to a convex hull of this 3D scan. This also has the added benefit of producing more accurate canopy volume estimations which are important for common precision viticulture management processes such as variable rate applications.

Keywords: ultrasound, array, vine yield, canopy estimation, smart agriculture, nondestructive, remote sensing

1. Introduction

The ability to accurately estimation grape yield is important because it allows viticulturist to plan, increase profitability, and improve the quality of the grapes produced. Yield estimation allows viticulturists to implement precision agriculture techniques including crop thinning, variable rate applications, and selective harvesting [1]. Traditionally, manual processes are used to estimate yield such as visual inspection and cutting and weighing grapes within a section of the vineyard [2]. However, these manual processes can be

*Corresponding author
Email addresses: 1badenparr@gmail.com (Baden Parr), M.Legg@massey.ac.nz (Mathew Legg), F.Alam@massey.ac.nz (Fakhrul Alam)

time consuming and the generally low number of samples taken can lead to inaccurate estimations. There is a need for an automated technique to accurately estimate grape yield.

Computer vision techniques have therefore been developed for automatically counting the number of grapes visible in camera images, and a high accuracy has been reported [3]. However, one limitation is that these techniques rely on being able to see the grapes. Errors in grape yield estimation occur where grapes are occluded by leaves or other grape bunches [4]. This has been addressed by assuming a certain percentage of grapes are occluded and compensating using a scaling factor [5]. However, this is not ideal and can lead to errors. Another approach is to remove leaves from the grape vines which could cause occlusions [6, 7]. However, this can be laborious unless specialised machinery is available. In addition, we understand that there are grape verities such as Gewürztraminer where foliage is normally not removed. Occlusion is perhaps the most significant unsolved issue for yield estimation using computer vision solutions.

One solution that has been suggested to address the issue of occlusion is microwave-based yield estimation [8]. The high frequency radio waves are able to propagate through foliage and reflect off the grape clusters behind. However, these devices are expensive and are not near commercial implementation. In this paper, we explore a previously unexplored alternative technology, ultrasound, for image through the leaves and detecting occluded grape bunches.

Within the field of precision viticulture, there have been several studies that have used ultrasound to map the outer leaf canopy shape for improved vineyard management. Gil et al. used three ultrasonic sensors to independently measure the distance to the vine foliage from spray nozzles positioned at different heights [9]. These transducers were positioned vertically in a line (tens of cm apart) and were operated independently to measure the distance to the foliage at three different heights. They were not used as an array. The closest distance reported by each sensor was used in real-time to control the application flow-rate from the nozzles. The benefit of this approach was verified by Llorens et al. who established that an average of 58% saving of application volume was obtainable [10]. In addition to variable rate application, independent scans taken over the growing season have been reported to have the potential to be an effective approach to monitoring vine vigour [11]. However, the effectiveness of these studies was limited by their use of ultrasonic transducers, which operated independently and not as arrays, to measure the distance to the outer surface of the foliage. These individual transducers have had a relatively wide beamwidth, and generally, the only information used from the reflected signal is the time of first echo from the foliage [11]. This results in low resolution imaging of the grapevine outer canopy and can give an overestimation of the canopy volume due to a few outer leaves sticking out [12]. Further work by Llorens et al. compared the same ultrasonic canopy measurements to a colocated 2D Light Detection and Ranging (LIDAR) scanner, a common alternative approach [13]. They found that the precise directionality of the laser distance measurements resulted in significant improvement in canopy surface estimation, albeit at the cost of a more complicated postprocessing procedure [14]. This highlights the utility that narrower beam-width

ultrasonic sensors may offer.

Recent work by Palleja et al. utilised four ultrasonic transducers to generate a volumetric estimation of a vine canopy using the signal envelope of multiple echoes [15, 16]. In a similar manner to Gil et al. [9], these transducers were arranged vertically in a line with each transducer being spaced 45 cm apart. They were not used as an array but as four transducers operating independently. However, the transducers employed had a wide beam pattern and therefore poor imaging resolution. For busy scenes, an independent ultrasonic transducer will be sensitive to multiple echoes from objects in a wide field of view. This is beneficial for applications such as a car reversing system where the system is only interested in the distance to the closest object. However, for an imaging system where one wants to image through leaves, using a single ultrasonic transducer will result in poor angular resolution. This is not desirable as it will make it hard to detect structure behind the closest leaf, see Figure 1a. Traditionally, one might increase the directionality of ultrasonic transmission by using transducers which are operational at high ultrasonic frequencies (several hundred kHz). However, we anticipate that this would come at the expense of reduced penetration through foliage and increased attenuation. These difficulties may explain why no previous studies have been found in the literature that have used ultrasound to image fruit occluded by leaves.

Arrays of ultrasonic transducers can be used to increase angular resolution [17, 18]. Figure 1b shows how an array of ultrasonic transducers can achieve a higher angular resolution compared with a single transducer. This significantly improves the potential for imaging structure behind the outer leaves. However, no previous study has been found which has used ultrasonic arrays to image any type of foliage apart from the authors' work with pasture in references [20, 19].

In this study, we present the first work where an ultrasonic array has been used to image grapes and foliage. To achieve an adequate angular resolution at lower ultrasonic frequencies (<60 kHz), we have utilised a novel air-coupled ultrasonic array developed by the authors [19, 20]. Another issue with using low ultrasonic frequencies is the low depth resolution due to the large wavelengths and ringing of the transducers [21]. This has been addressed in this work using coded waveforms, cross-correlation, and operating away from the transducers' resonant frequency. Ultrasonic arrays and coded waveforms have not been used before in precision viticulture.

The high spatial and depth resolution from the array allowed the echoes from grapes and leaves to be separated. However, the ultrasonic echoes from leaves and grapes appeared to be identical. This was addressed by making multiple ultrasonic measurements at the same location while lightly agitating the leaves with a fan directed at the measurement area. Since the leaves moved while the heavier grapes remained stationary, the mean and variance the ultrasonic measurements could be used to identify the grape bunch.

Initially, imaging was performed with the array focused in the far-field. Work was then undertaken to investigate the improvement in imaging resolution using near-field focusing of the array. This includes a novel technique to compensate the cross-correlation for near-field defocusing of the transmitted signal.

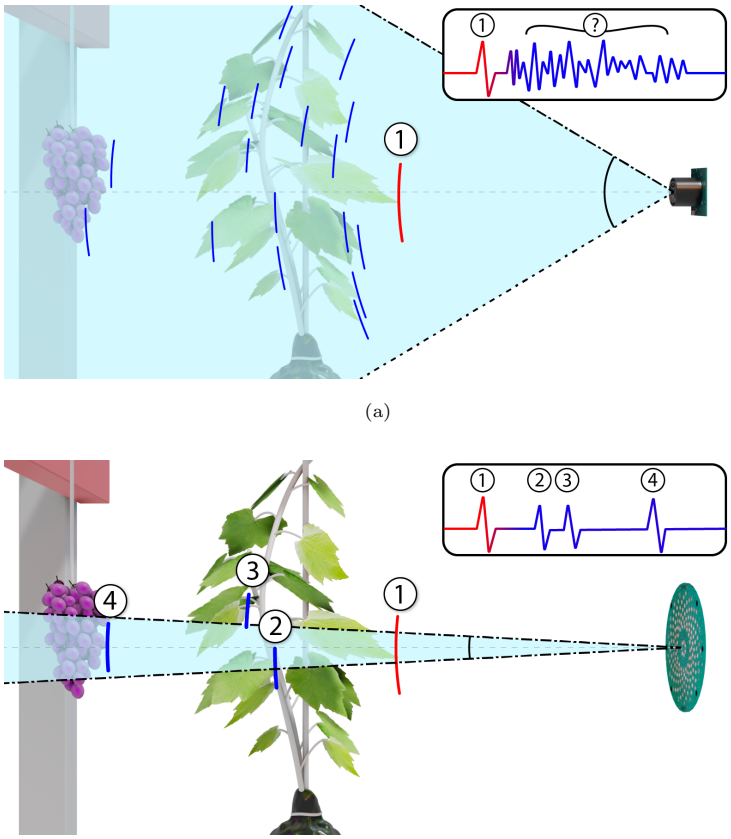


Figure 1: A single ultrasonic transducer (not an array) with a wide beamwidth can struggle to image objects behind the front leaves, as illustrates in diagram (**a**). In contrast, diagram (**b**) illustrates how an ultrasonic array such as used in this work (see Figure 2 for a photo) with a narrow beamwidth can provide improved ability to image at multiple depths behind the front leaves.

The improved spatial resolution in the resulting volumetric scans will be a benefit for precision viticulture management processes such as variable rate applications where an accurate understanding of the vine canopy is vital.

This paper has the following significant contributions to knowledge. It is the first work to use an air-coupled ultrasonic phased array and coded waveforms for the purpose of analysing vine canopies. It is also the first study to investigate if it is possible to use ultrasound to image through leaves, to detect fruit located behind leaves, and to differentiate echoes that come from leaves through agitation. In addition, we present a new technique for improving the resolution of the array based cross-correlation for near-field echoes. This approach simulates the effect of focusing the transmission of the array at any desired depth in postprocessing. This eliminates the need for the complex electronics required for focusing the array's transmission to a desired scan depth. Some preliminary results of this work were presented in the conference paper [22].

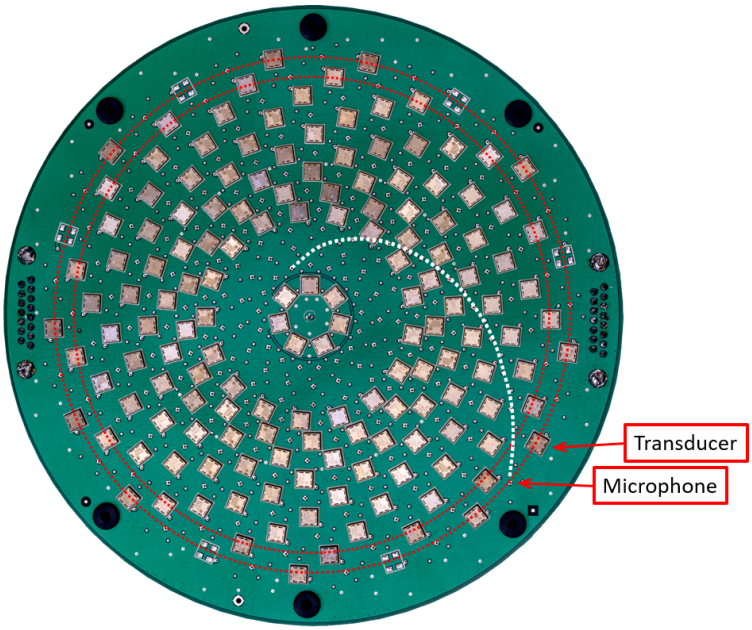


Figure 2: Photo of the ultrasonic array's main PCB. The transducers (silver squares) and microphones (located behind holes) are arranged in a multiarm spiral pattern forming rings. One of the microphone spirals arms is illustrated by a white dashed line while the two outer rings of microphones and transducers are shown as red dashed lines.

The paper is organised as follows. Section 2 introduces the ultrasonic array hardware and measurement parameters used in this work. The experimental setup and measurement procedure are described in Section 3. The signal processing applied to the array data for imaging grapes is then presented in Section 4. Sections 5 and 6 provide results for the array focused in the far-field and near-field respectively. Finally, in Section 7, we end the paper with some final points and a discussion about future directions that can be taken.

2. Ultrasonic Array

Figure 2 shows the ultrasonic array thas has been used in this work. This was custom designed and built by the authors for precision agriculture requirements. A full description of this array is given in reference [19]. It has optimised spiral arrays of 160 ultrasonic transducers and 204 microphones, which are arranged into rings. The transducers are surface mounted to the front of the array PCB. In contrast, the MEMS microphones (which can operate at ultrasonic frequencies) are surface mounted to the back of the PCB with holes passing though the PCB to allow the acoustic signal to be measured. The radius of the transducer and microphone rings are given in Table 1.

The microphone array had 12 independent rings of microphones. All the microphones in a ring were connected in parallel and then captured by one of 12 simultaneous sampling Analogue to Digital Converter (ADC) channels of a Data Translation DT9836 module [23], refer to Figure 12b in reference [19]. A sampling

Table 1: Radii of microphone and transducer rings.

Ring Number	Microphone (mm)	Transducer (mm)
1	15.0	9.0
2	20.3	31.0
3	25.7	36.4
4	31.0	41.8
5	36.4	47.3
6	41.8	52.8
7	47.3	58.3
8	52.8	63.8
9	58.3	69.4
10	63.8	75.0
11	69.4	
12	75.0	

rate of 225 kHz and a resolution of 16 bits were used. Note that since all 12 microphone ring channels were saved to file, it was possible to dynamically change the focus distance of the reception in postprocessing using beamforming.

The transducers used in the array were surface mount air-coupled transducers which had a resonance frequency of 40 kHz and a frequency response which dropped from this peak by about 20 dB at 25 kHz and 60 kHz on either side. The measured frequency response can be seen in Figure 3. Although the transmission gain is highest around 40 kHz, the transducers have a tenancy to ring at this resonance frequency, which is undesirable if cross-correlation is being used to improved depth resolution. We therefore operate them at frequency ranges on either side of the resonant peak (e.g. 20–35 kHz and 45–60 kHz). The transducers were arranged in 10 rings. The DT9836 board's two Digital to Analogue (DAC) channels were used to drive the 10 rings (half of the rings for each DAC channel) through two power amplifiers, refer to Figure 12a in reference [19]. These had an output sampling rate of 500 kHz and resolution of 16 bit and were synced with the ADC channels. Data acquisition software was written in MATLAB to transmit the signal and capture the resulting echoes using the DT9836 board.

The same excitation signal (a linear chirp) was applied to all the transducers. Since the array was planar (on a flat PCB), this meant that it was effectively using far-field beamforming with the transmission focused at a point in front of the array at infinity. Near-field focusing was not possible for transmission since we did not have a separate DAC channel controlling each transducer ring. Figure 4 shows the measured

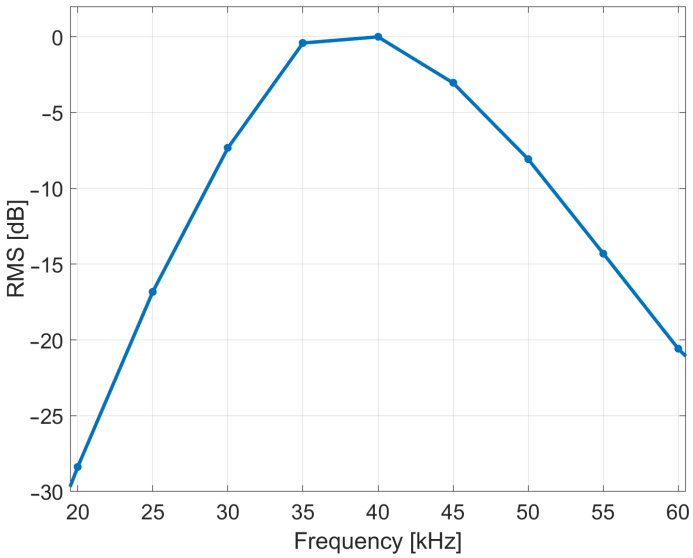


Figure 3: The surface mount transducers frequency response that was measured using the microphone model that was used in the array.

combined transmit/receive beam pattern of the array when the array is focused at infinity. This shows a full beamwidth of 3.3° and a dynamic range of up to 33 dB. Please refer to reference [19] for details on how this beam pattern was obtained. The array had a dead-zone of about 500 mm where the signal measured by the receiver channels was dominated by the vibrations caused by the ultrasonic transmission. Objects closer than this were hard to detect.

Air-coupled transducers generally achieve a high gain at the expense of ringing at the resonant frequency of the transducer. As a result, digital codes such as Barker Codes or Maximum Length Sequence (MLS) with sharp transitions can cause ringing and may not be reproduced correctly by the transducers. In contrast, the lack of sharp temporal transitions for a chirp waveform means that it is less prone to exciting ringing of the transducer compared with some other waveforms.

The transmit linear chirp signal applied to the transducers can be described by

$$\boldsymbol{y}[n] = \boldsymbol{W}[n] \cdot \sin\left(2\pi\left[f_0 \;\; t[n] + \frac{\beta}{2}\,\frac{\boldsymbol{t}[n]^2}{\tau}\right]\right), \tag{1}$$

where $\boldsymbol{t}[n]$ is the time of the n_{th} transmit sample, f_0 is the start frequency at $t = 0$, β is the bandwidth, τ is the pulse duration, and $\boldsymbol{W}$ is a Hamming window [24].

A chirp excitation signal was chosen as it can be used with cross-correlation to improve the depth resolution. After testing, a linear chirp with a duration of 1.5 ms and bandwidth of 45 to 60 kHz was chosen. This transmitted signal was verified through independent recording using a calibrated microphone (GRAS 46BF-1 1/4 inch). The signal time and frequency domain representations can be seen in Figure 5. The small time delay is due to the separation between the transducer and microphone. The 1.5 ms duration chirp

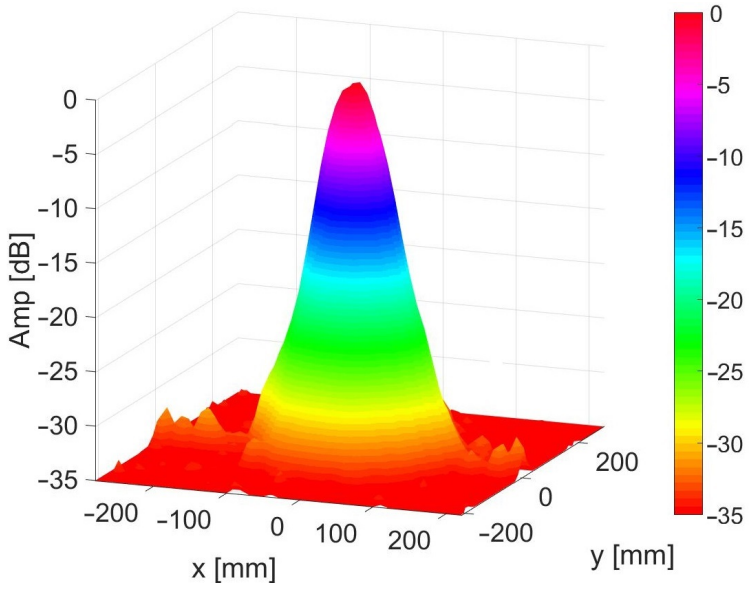
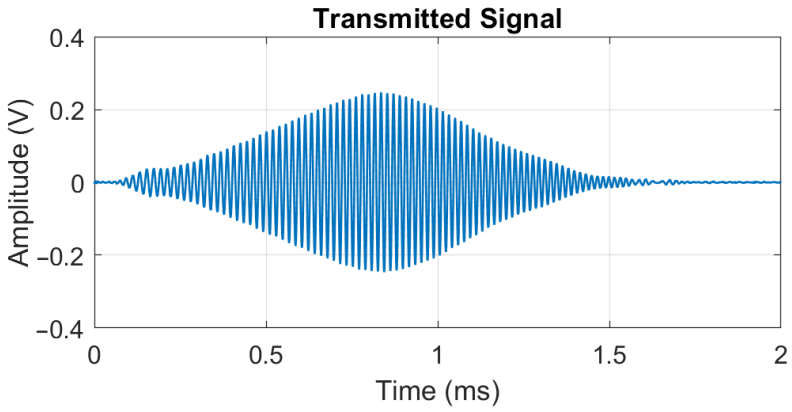
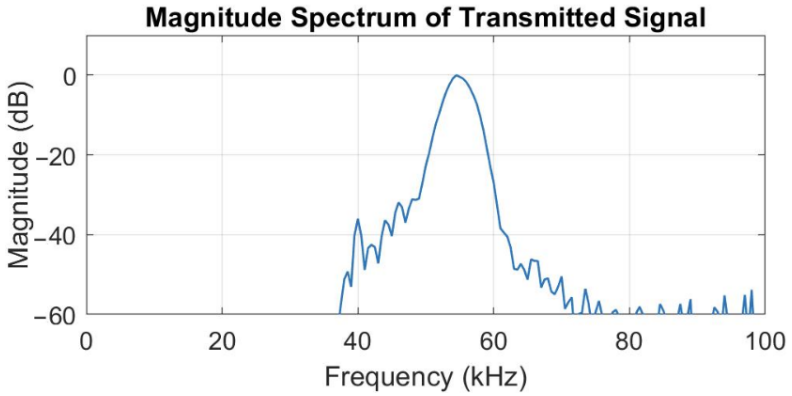


Figure 4: Plot of the measured array beam pattern (combining transmission and reception) for a 35 kHz sine wave using a 40 mm diameter reflector at 805 mm distance from the array.

appeared to provide improved cross-correlation resolution compared to a shorter duration chirp. The frequency bandwidth was chosen based on the frequency response of the transducers, which have a usable frequency range between 25 kHz and 60 kHz and a resonant frequency of 40 kHz, see Figure 3. To avoid ringing at this resonant frequency, the chirp used a frequency range from 45 to 60 kHz. It was also felt that this frequency range gave slightly better depth resolution than the 25–35 kHz range due to the smaller wavelength. The attenuation experienced by the ultrasound as it travels through the air can be calculated using the atmospheric absorption model given by International Standard ISO 9613-1:1993 [25]. It can be shown that at a standard atmospheric pressure, a temperature of 20°C and 50% humidity at 60 kHz this is about 1.98 dB/m. For practical operation in a vineyard, the width of the rows limits the operating distance to about 1 meter. Over this distance, the attenuation is negligible.



(a)



(b)

Figure 5: Plots showing (**a**) the ultrasonic chirp transmit signal as recorded by a calibrated microphone (GRAS 46BF-1 1/4 inch), and (**b**) its corresponding frequency domain representation. The small peak seen at 40 kHz is a result of ringing at the transducers' fundamental frequency.

3. Experimental Set-up and Procedure

The ability of the array for detecting grapes was evaluated using a 2D Computer Numerical Controlled (CNC) gantry system. This CNC had a range of motion of 1.4 × 1.4 m and a resolution of 0.025 mm. Ideally the array would have been mounted to the CNC machine. However, the array was originally designed for operation from a farm vehicle and was too heavy in its current mounting. Instead a grape vine was mounted directly to the CNC machine. The grapes were fixed to 3 mm rods. This was done to reduce the amount of movement when the CNC was moving and to minimise reflections from this support. The vine was mounted to a bamboo pole and its roots were surrounded by a plastic bag with most of the soil removed to reduce weight. Refer to Figure 6 for a photo of the setup.

Initially acoustic foam had been placed behind the CNC to dampen echoes from the wall behind, as shown in Figure 6. However, subsequent measurements were made with the foam removed and no noticeable

Figure 6: Photo showing the experimental setup with the grapes located behind a vine that is attached to a Computer Numerical Controlled (CNC) machine.

difference in measurement performance was observed. This was expected as the array has a highly directional beamwidth and as a result is very insensitive to reflections outside its field of view, as shown in Figure 4. This shows that in a field environment such precautions would not be necessary.

The experimental setup and measurement scan volume are illustrated in Figure 7. The ultrasonic transducer was positioned facing the CNC machine at a distance of 1100 mm in front of the grapes. Measurements were made over a 460 $\times$ 400 mm wide grid with a spatial separation between ultrasonic measurements of 20 and 50 mm in the x and y axis respectively. This gave 216 measurement points. Between each ultrasonic measurement, the CNC was paused 3 seconds to allow time for the vine and grapes to stop moving before ultrasonic measurements were made. This measurement procedure was repeated for each of the types of scans described below.

It was anticipated that it would be challenging to differentiate echoes from leaves from that of grape bunches. To try to address this, ultrasonic measurements were therefore made with a fan lightly agitating the leaves, while the heavier grape bunches remained stationary. This agitation could be achieved in the field using a fan or even possibly utilising naturally occurring wind.

The following sets of ultrasonic measurements were therefore made for (a) grapes only with no vine present, (b) both grapes and vine with no fan, and (c) grapes and vine with the fan operating. The fan was was pointed in front of the array and used to lightly agitate the vine leaves. Using a handheld anemometer, the wind-speed at the location of the vine-foliage was measured to be 2.5 m/s. More work is needed in the future to investigate the relationship between air-speed and the resulting agitation performance.

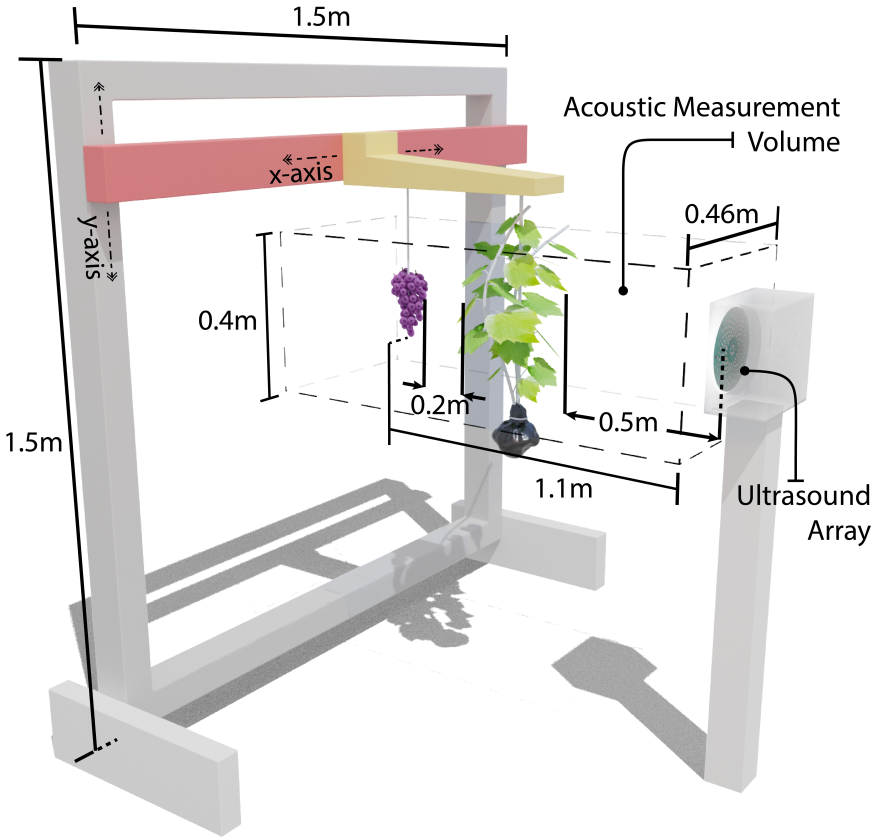
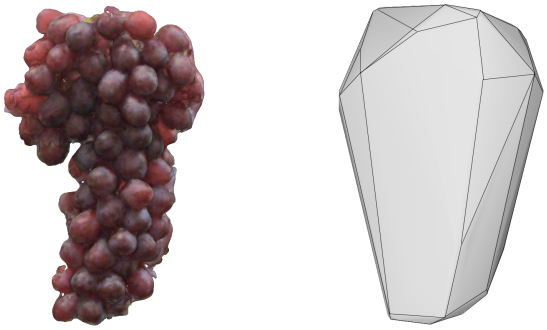


Figure 7: Diagram of the experimental setup showing the grapevine and grape bunch suspended from the CNC machine in front of the ultrasonic array and volume of area where ultrasonic measurements were performed.

3.1. Measurement of Grape Volume Using Photogrammetry

In Sections 5 and 6, the ultrasonic measurements are processed to provide an estimate of the volume of the grapes. To provide a comparison (ground truth), the volume of the grapes needed to be measured using an alternative technique. A photogrammetry process was therefore used to construct an accurate 3D scan of the grape cluster. This was achieved by using Agisoft Metashape Professional v1.5.2 to process 30 images captured by a Sony A6300 covering the grape cluster from all sides. The resulting scan can be seen in conjunction with a convex hull approximation in Figure 8.

We have used a convex hull as it offers a representation closer in likeness to the results of this acoustic scan, in that, the concave details of the individual grapes are removed. The convex hull was computed using the convex hull tool in Meshlab 2020.07. The volume of the 3D scan and convex hull are given in Table 2.



(a) (b)

Figure 8: Renders of the 3D scan (**a**) of the grape cluster constructed using photogrammetry and the corresponding convex hull (**b**) created in MeshLab 2020.07.

Table 2: Grape volume measured using photogrammetry.

	3D Scan	**Convex Hull**
Grape volume [ml]	580	1200

4. Processing Array Data

4.1. Beamforming to Improve Spatial Resolution

An array of ultrasonic sensors can achieve much higher resolution than can be achieved from a single ultrasonic sensor [26, 27]. For reception (RX), this was achieved by combining the 12 microphone receiver channel signals into a single channel of data using beamforming.

The echoes from objects were captured by the $M = 12$ microphone ring channels. The record duration was 20 ms which corresponds to N = 4500 samples and a maximum resolvable depth of roughly 3.4 m. The microphone data was stored as a $[N \times M]$ matrix $\boldsymbol{x}$, where the m_{th} column corresponds to the data for the m_{th} microphone ring and is expressed as $\boldsymbol{x}_m$.

Figure 9 shows a CAD diagram of the ultrasonic array PCB with two of the microphone rings shown as circles. The time delays required to focus the array in the near-field at a point z in front of the array are also illustrated. The reception of the array can be focused at a desired distance z along a line normal to the centre of the array using beamforming, see Figure 9. To achieve this, a delay can be calculated for each microphone ring using

$$\Delta t_m(z) = \frac{\sqrt{r_m^2 + z^2} - z}{c}, \tag{2}$$

where r_m is the radius of the m_{th} microphone channel ring given Table 1, and c is the speed of sound in air. We can convert this delay to an integer number of samples using

$$\Delta n(z) = \text{round}\{\Delta t_m(z) \times f_s\}, \tag{3}$$

where f_s is sampling rate.

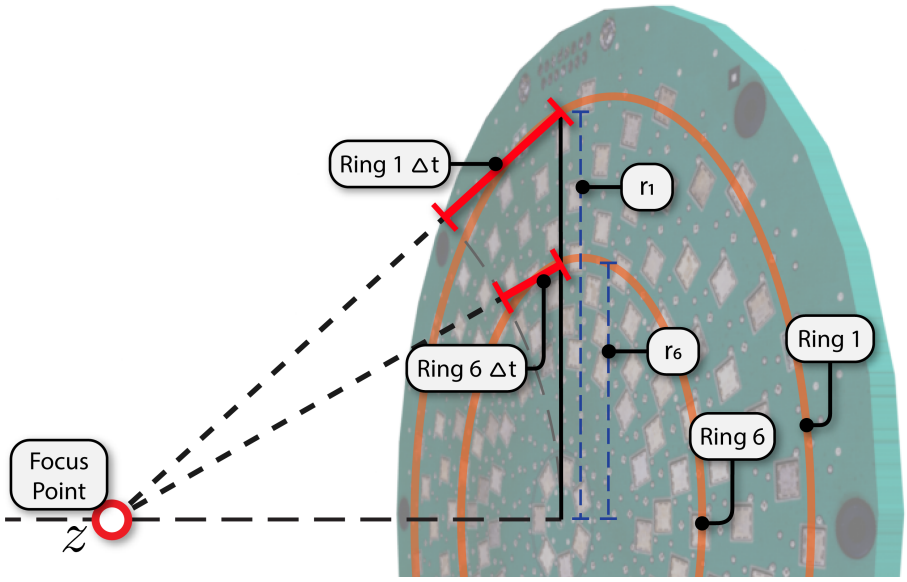
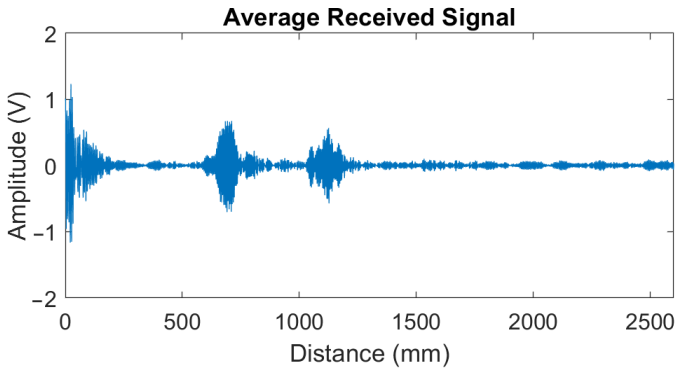


Figure 9: The difference in path length and hence time delay Δt is illustrating in this diagram for two of the array rings when focusing in the near-field at a distance z in front of the array.

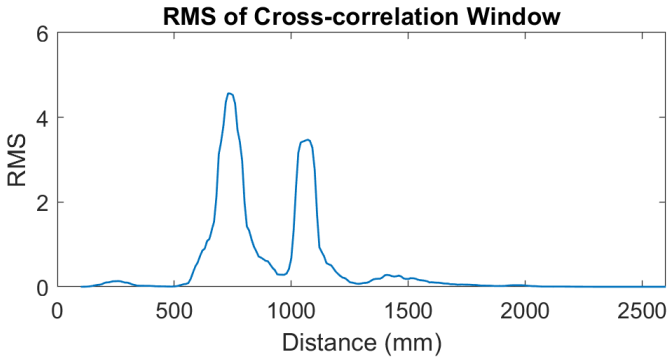
The $[N \times M]$ microphone channel data matrix $\boldsymbol{x}$ can be converted to a $[N \times 1]$ array of data $\bar{\boldsymbol{x}}$ which is focused at a distance z using delay and sum (time domain) beamforming

$$\bar{\boldsymbol{x}}[n] = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{x}_m[n - \Delta n(z)], \tag{4}$$

where n is the sample index. See Figure 10a for an example of this summed signal. Note that as z becomes large the beamforming delays Δn go to zero. From Equation (4), we can therefore see that averaging all 12 microphone channels (no delays) focuses the array at infinity. This will be referred to here as far-field beamforming. A 40 kHz notch filter was then applied to the resulting signal $\bar{\boldsymbol{x}}$ to remove the ringing at the transducers resonance frequency.



(a)



(b)

Figure 10: Example plots showing (**a**) the received signal with all the microphone rings averaged, and (**b**) the corresponding windowed RMS representation of the cross-correlated signal.

A problem with using time domain beamforming is that the delays that can be applied must be a multiple of the sampling interval, see Equation (3). This can mean that unless the sampling rate is high, the focus may not be accurate. An alternative technique is to use frequency domain beamforming since this does not have this quantisation issue and can therefore be more accurate. Frequency domain beamforming can be achieved by shifting the individual microphone channels using

$$\boldsymbol{X_{sm}}\left(\omega, \Delta t_m(z)\right) \; = \boldsymbol{X}_m(\omega)\; e^{\{-i\;\omega\;\Delta t_m(z)\}}, \tag{5}$$

where X is the complex discrete Fourier transform of the recorded signal and ω is the angular frequencies [28]. The phase shifted signal can then be converted back to the time domain using the inverse Fourier transform and summed into a single beamformed channel $\bar{\boldsymbol{x}}$.

There were 20 recordings made at each measurement location of the CNC, giving 20 sets of $\bar{\boldsymbol{x}}$ vectors. The average, $\boldsymbol{\mu}$ and variance, $\boldsymbol{\sigma}^2$ of these were then calculated element wise for each sample resulting in $[N \times 1]$ average and variance vectors.

The beamformed signal will contain peaks corresponding to echoes from reflectors in front of the array. The distance to the reflectors can be obtained by converting the time t_n when an echo peak occurs in the signal to a distance using

$$d = \frac{t_n}{2}\, c, \tag{6}$$

where c is the speed of sound. Note that the division of the time by 2 in this equation is due the fact that the echo signal has to travel twice the distance from the array to the object. The speed of sound in air can be approximated as

$$c = c_o \sqrt{1 + \frac{T}{273}} \tag{7}$$

where T is the temperature in degrees Celsius and $c_o = 331.5$ m/s [29]. The ambient temperature was measured using a temperature sensor included as part of the hardware. It was found to be 23 ± 0.25 °C giving a speed of sound of 344.8 ± 0.15 m/s. In an outdoors environment, the ambient temperature would be expected to fluctuate more. This would require the air temperature to be monitored closely to allow real-time compensation for the speed of sound on distance measurements.

4.2. Cross-correlation to Improve Depth Resolution

The temporal/depth resolution of the system was improved using cross-correlation [30]. The cross-correlation was calculated using

$$\boldsymbol{r}_{hx}[n] = \sum_{k=-\infty}^{\infty} \boldsymbol{\mu}[k]\boldsymbol{h}[k-n], \tag{8}$$

where $\boldsymbol{h}[n]$ is a filtered version of the signal $\boldsymbol{y}[n]$ applied to the transducers [31]. This filtered the signal to simulate the frequency response of the transducers which is shown in Figure 3. An example of the result of this process can be seen in Figure 10b where the two resolved echoes correspond to the leaves and grapes.

4.3. Correction for the Array's Transmission Being out of Focus

The above cross-correlation technique assumes that the ultrasonic echoes from a point source located directly in front of the array will result (after beamforming and averaging of the received signal) in a signal $\boldsymbol{\mu}$ that is a scaled and delayed version of the transmit signal $\boldsymbol{h}$. However, for a planar array, this is only true if the array is correctly focused (correct beamforming time delays are applied for each transmission and reception array channel). Incorrect focusing of the array will result in signal $\boldsymbol{\mu}$ being received from a point reflector that is distorted and not a scaled version of the transmit signal $\boldsymbol{h}$. This distortion will cause reduced efficiency/errors in the cross-correlation technique.

The reception of the array is able to be focused in postprocessing for any desired distance from the array since each of the 12 microphone rings was sampled using an independent ADC channel. However, for the transmission, this was not possible since the transducer rings were wired in parallel. Even, if the transmission could have been focused (if they had an independent DAC and power amplifier per transducer ring), it would

have only been possible to focus at one distance per transmission. Unlike reception, transmission focusing cannot be done in postprocessing. This means that multiple transmissions would be required to allow focusing at a range of distances in the scan volume.

To overcome these issues, a technique was therefore developed to correct for this near-field distortion effect of the transmission in post process. Rather than using the signal $\boldsymbol{y}(t)$ that was applied to the transducers for cross-correlation, a new distorted version of this transmit signal was simulated using

$$\bar{\boldsymbol{y}}[n] = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{y}[n - dn(z)]. \tag{9}$$

This distorted simulated signal was then bandpass filtered by the frequency response of the transducers to give $\boldsymbol{h}[n]$ and used with Equation (8), the beamformed and averaged reception signal $\boldsymbol{\mu}$, to give the cross-correlation $\boldsymbol{r}_{rh}[n]$ for any desired imaging distance z. For each scan, this process was repeated for a range of distances. We have not been able to find this technique being used before in the literature.

4.4. Estimating Volumes of Scattering Objects

The cross-correlation signal could be plotted as function of distance by converting sample times to distance using Equation (6). A sliding window with a width of 26 samples and a 50% overlap was then used to convert the cross-correlation data to an array of RMS values, where the distance separation between RMS values was 10 mm. This RMS windowing technique was implemented over the scan volume of size 460 × 400 × 900 mm, as shown in Figure 11. Within this volume, 19,224 scan points were defined by dividing the volume up respectively into 24 × 9 × 89 uniformly spaced points.

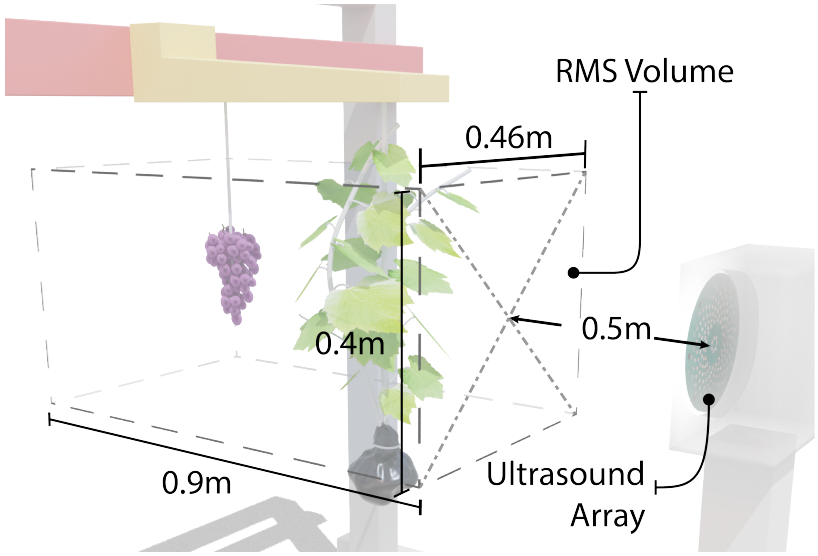


Figure 11: Diagram showing the scan volume used for the RMS processing for grape and leaf detection.

Isosurfaces are used to visualise the computed volumes using a threshold of 10% of the maximum RMS for each scan. This results in a 3D surface representation of the volume that encompasses all points that have a value at least 10% of the maximum RMS recorded. If this threshold were 0% then the isosurface would represent the entire measurement volume. This threshold was empirically determined to best demonstrate the response of the system. The numerical volume of each isosurface can be naively determined by treating the grid points as voxel cuboids and counting those that are over this threshold within a given region. Each voxel in this measurement corresponds to $20 \times 50 \times 10$ mm $= 10$ mL. Other techniques for measuring the volume could be investigated in the future such as mean-shift or k-means clustering [32].

5. Results for Far-Field Focusing of the Array

Scans were first made without agitating the grapes with a fan and averaging of repeated samples. An example of a resulting RMS isosurface can be seen in Figure 12. This shows two volumes corresponding to the leaves with the grapes behind. This plot shows that the grapes can be detected behind leaves.

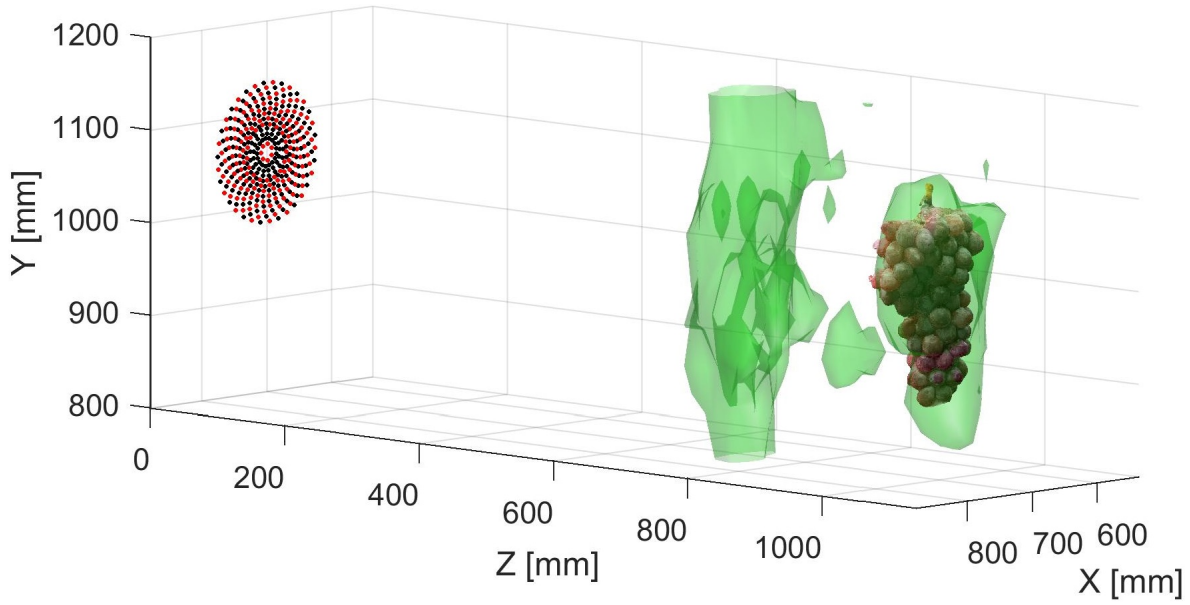


Figure 12: An plot showing an RMS isosurface plot for scans of a grape bunch (at about $z = 1100$ mm) and leaves (at about $z = 700$ mm), where 20 averages were made of the measurement at each position. The 3D photogrammetry scan of the grapes and the microphone (black dots) and transducer (red dots) arrays have been overlaid. It can be seen that the volume of the leaves is significant.

5.1. Differentiation of Leaves and Grapes

The measurements described above showed that the grapes and leaves could be detected using ultrasound. However, this technique did not allow one to identify if the reflections were coming from leaves or grapes. We

Table 3: Grape and foliage volume estimates using ultrasonic far-field array focusing.

	Static (fan off)	**Averaged (fan on)**	**Variance filtered (fan on)**
Grape volume [ml]	1660	1870	3500
Leaf volume [ml]	8510	3890	870

believed that agitating the leaves with a fan might allow this to be achieved. For each position of the CNC machine, 20 recordings were made. The microphone signals from these recordings were averaged (equivalent to far-field beamforming) and the variance obtained.

Figure 13 shows the resulting isosurface plot after the ultrasonic echo signal had been filtered using average and variance. The movement of the leaves resulted in an increased variance for the the leaves compared to that of the grapes. This technique was able to remove almost all of the signal from the leaves and identify the grapes. A further filter could be added to remove isolated smaller isosurfaces that had an area too small to be expected to be a grapes. Table 3 compares the estimated volumes of the grapes and leaves using these techniques.

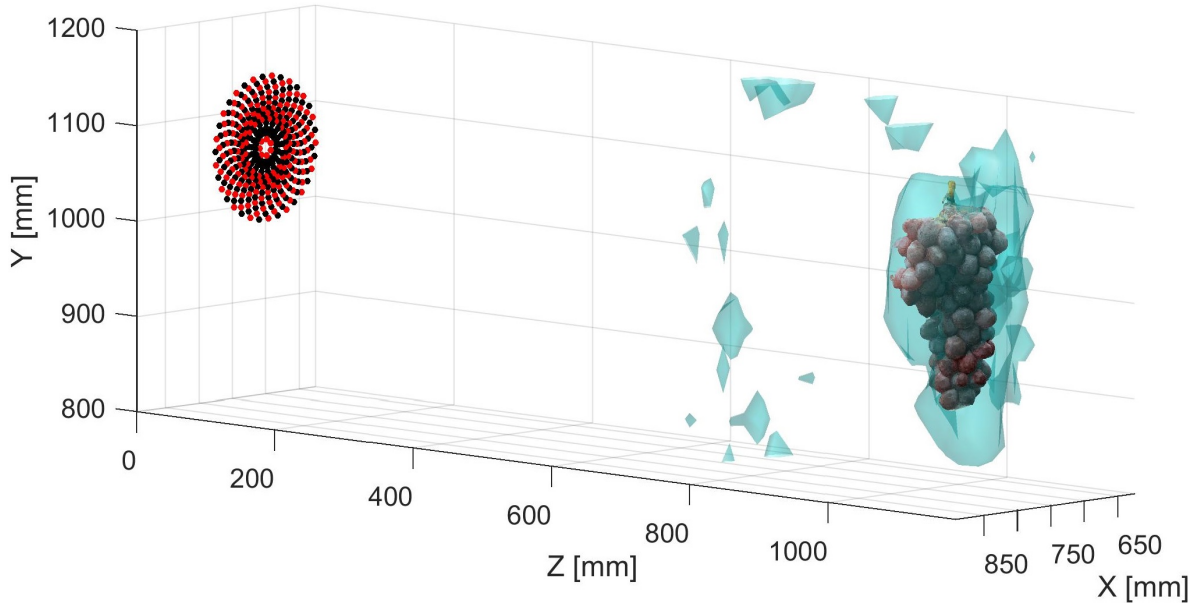


Figure 13: The isosurface plot was achieved using a fan to agitate the leaves and performing filtering of the signal using the average and variance for 20 recordings. When comparing this plot with Figure 12, one can see that this technique allowed one to mainly remove the echoes from the leaves.

6. Results for Near-field Focusing of the Array

The results shown this far present a potential process for the identification of grape clusters in the presence of foliage. However, while promising, the results suggest that more resolution and detail of the canopy can be obtained if the acoustic array had a narrower beamwidth. Although the inherent far-field beamwidth of the array is very narrow, it still diverges at roughly 3.3 degrees. At a distance of 1 m, this equates to a circular cross section of around 58 mm, making it difficult to distinguish between tightly packed objects. Decreasing this beamwidth further would improve the array's ability to reject reflected sound from nearby objects. Near-field focusing of the array could help improve imaging resolution and hence provide more accurate representation of the scene resulting in a better understanding of the true canopy volume.

As discussed in Section 4.1, we can achieve near-field focusing of the array using beamforming of the microphone/receiver signal (RX beamforming). With this approach, the microphone receiver array can be focused at a particular distance from the centre of the array, increasing sensitivity at that point and reducing sensitivity to surrounding points. It will also minimise distortion of the signal, which will improve cross-correlation performance. This focusing can be achieved by calculating the phase difference of arrival to each microphone from a sound wave reflected off an object at the focus distance. A corresponding phase shift is then applied to each microphone channel's recording. The simulated beam patterns shown in Figure 14 indicate that focusing the array in this way could improve the angular resolution substantially. These beam patterns were generated by simulating the sound propagation from each transducer in the array to a reflector situated at a perpendicular distance of 700 mm from the face of the array. The received signal after processing is compared to the transmitted signal using the maximum cross-correlation as discussed in Section 4.2. The resulting maximum correlation for each x position is shown in power form, normalised to 0 dB. The simulation shows a significant reduction in -3 dB beamwidth, from 44 mm to 17 mm, and reduced sidelobes when the correct focus obtained with near-field beamforming is used.

We could extend this process further by applying the same technique to the transmitted signal. As the transducers also have a significant spatial separation, a synchronously transmitted waveform from each ring of transducers, will reach a particular focus distance at slightly different times. This will cause the apparent signal at that point to become distorted. Traditionally beamforming of transmission signal (TX beamforming) would be performed before transmission to compensate for these delays and ensure the signal reaches its target distance undistorted. The result of performing this TX beamforming is shown simulated in Figure 14. It shows a marked improvement in -3dB beamwidth, from 44 mm to 16 mm when using both RX and TX beamforming. Furthermore, sidelobes suppression is significantly improved, showing a 13 dB improvement over just using RX beamforming.

RX beamforming provides a significant improvement to the arrays performance and can be applied to a single recording for all distances in postprocessing. Unfortunately, traditional TX beamforming requires

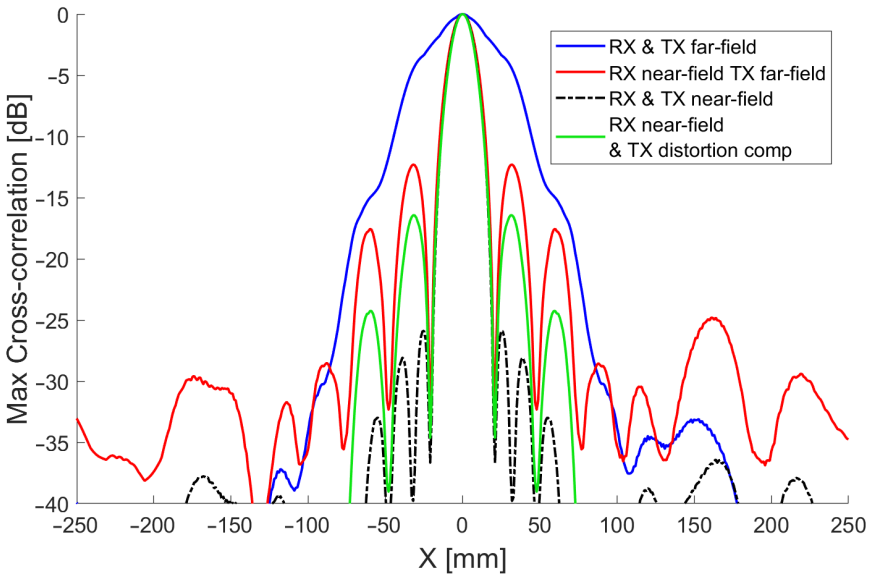


Figure 14: Simulated 60 kHz cross-correlation beam patterns for echoes from a point source reflector located 700 mm from the array employing far-field and near-field beamforming for reception (RX) and transmission (TX). Shown in green is the combined RX near-field and TX distortion compensated beam pattern that was used in this study.

multiple transmissions to cover the entire depth range which makes it largely impractical for our situation. These unique transmissions would take a considerable amount of time to perform and increases the complexity of deploying a real-time system for use in vineyards. Furthermore, it removes the ability to evaluate different focus distances after measurements are conducted. In some situations, it may be beneficial to change the z-axis resolution to get a more detailed view of the scene. Additionally, TX-focusing requires additional hardware in terms of an independent DAC and power amplifier per transmission ring.

To work around these limitations, in Section 4.3, we introduced a novel technique to compensate the cross-correlation for the distortion that affects a transmitted signal when it is not correctly focused. This has the benefit of being computed after capture during postprocessing, in conjunction with RX beamforming, allowing for optimal results at all distances with a single scan. Equation (9) from Section 4.2 describes how the distorted signal at a desired depth can be calculated to then enhance cross-correlation performance. The simulated performance of this technique can be seen in Figure 14. The process results in a substantial improvement over just using RX beamforming. Sidelobes see a further 4 dB of suppression and the -3 dB beamwidth is reduced from 18 to 16 mm. These improvements translate to more granular resolution in the 3D volumetric scans of grape vines. The narrower beamwidth should allow more detail to be captured of the vines and the reduced sidelobes will reduce susceptibility to multipath interference from nearby foliage and other reflectors.

If we repeat the process used to generate an RMS volume as discussed in Section 4.4, we can compute a

Table 4: Grape and foliage volume estimates obtained using near-field RX and TX focusing of the array for different types of scans.

	Averaged	Variance Filtered
Grape volume [ml]	1230	1320
Leaf volume [ml]	3470	280

comparable volumetric representation using the improved near-field beamforming technique. The resulting RMS volume shown in Figure 15 is presented as an isosurface with the threshold set to 10% of the maximum RMS. A direct comparison can be made to the unfocused scan seen in Figure 12. As can be seen, there is a significant increase in the level of detail in the 3D volume. The volume is less globular and more defined. Increasing observable detail of the structure of the vine canopy could lead to improved vineyard management through more precise knowledge about foliage density and crop loading.

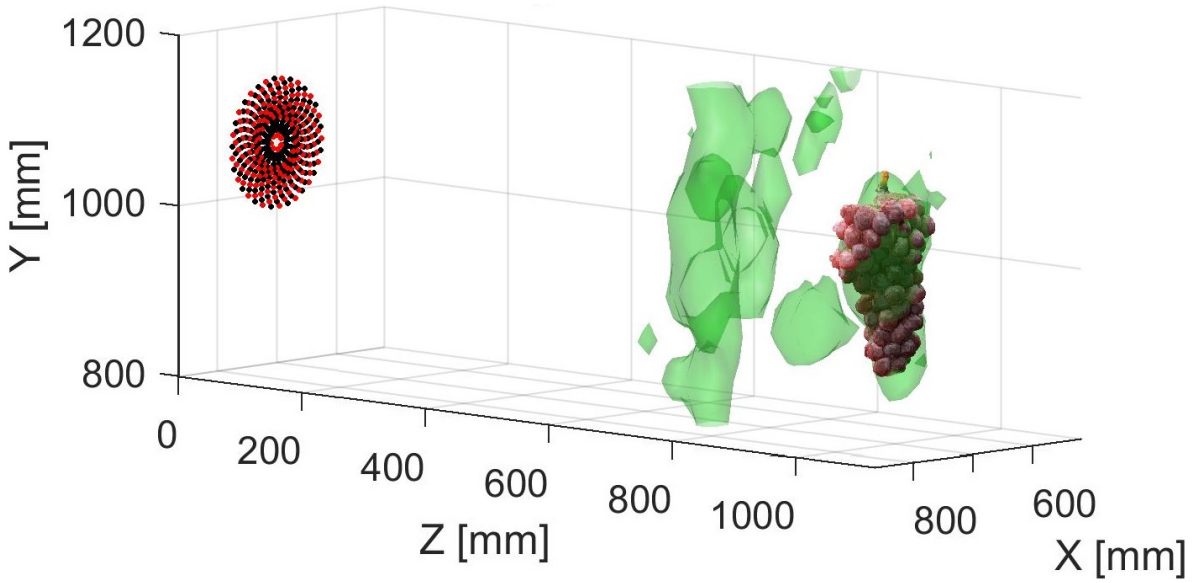


Figure 15: Isosurface visualisation of near-field RX beamformed and TX distortion compensated acoustic scan with averaging of 20 recordings at each scan point. The additional level of detail can be clearly seen in the focused RMS volume.

The reflections from the leaves can be mitigated in the focused scans using the technique described in Section 5.1. A fan was used to agitate the leaves. Filtering was performed using averaging and variance. Figure 16 shows the resulting isosurface plot. Table 4 compares the grape and foliage volumes obtained using the near-field focused techniques.

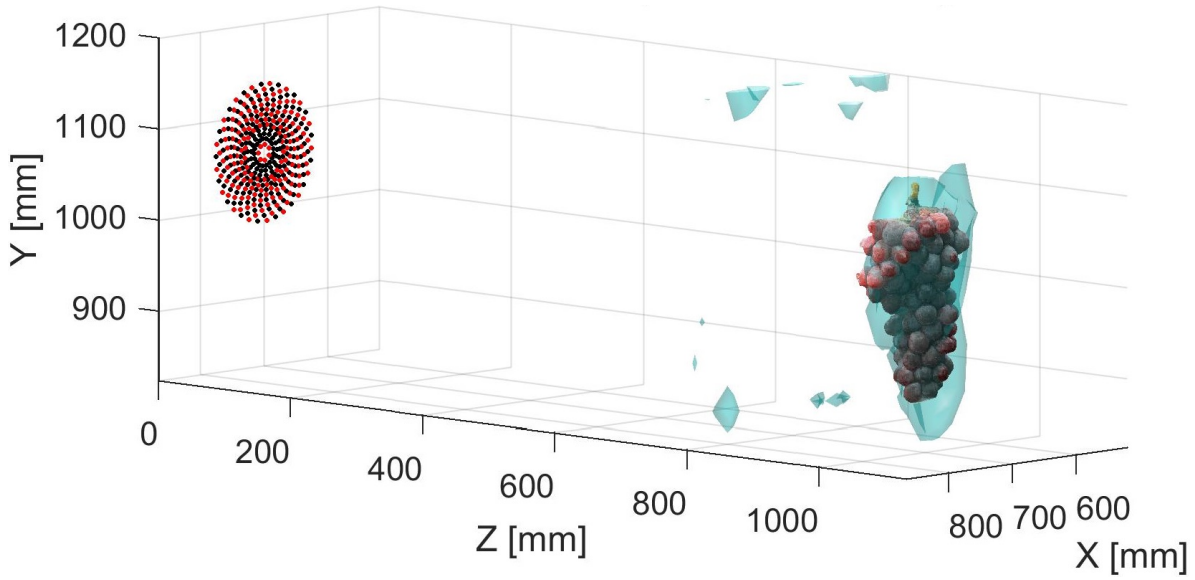


Figure 16: Isosurface visualisation of near-field RX beamformed and TX distortion compensated acoustic scan where echoes from leaves had been mitigated using a fan to agitate the leaves and performing filtering using the average and variance of 20 recordings at each scan point.

7. Conclusions and Future Work

This paper presents a novel approach for the detection of grape clusters which are occluded by foliage using an ultrasonic array. It utilises a low frequency ultrasonic chirp transmitted from a highly directional acoustic array. This is the first time that an ultrasonic phased array has been used to analyse canopy structures and the first time ultrasound has been used to visualise grape clusters. The results show that it is possible for low frequency ultrasound to penetrate through leaves and generate echoes from the grapes behind. In addition, the echoes from grapes and leaves can be distinguished by agitating the leaves using a fan and using the variance of multiple recordings as a filter.

We further demonstrate how increased detail in the acoustic volumes can be achieved through near-field focusing the reception of the array using beamforming and cross-correlation defocusing correction techniques. This significantly reduces the beamwidth and increases directionality of the array. The increased level of detail has direct benefit for more accurate canopy estimation and as a result, improved precision viticulture practices.

Improved spatial and depth resolution would also be expected to reduce the overestimation in volume measurement obtained using the ultrasonic measurement. Table 5 compares the percentage overestimation in volume obtained in Tables 3 and 4 using the ultrasonic methods compared to the volumes of the the 3D photogrammetric scan and the convex hull given in Table 2 which was obtained using photogrammetry. Here we can see that the use of near-field focusing techniques with averaging reduced the ultrasonic measured

Table 5: Percentage overestimation of the measured grape volume obtained using near-field and far-field focusing compared to the grape volume obtained using photogrammetry 3D scan (580 mL) and convex hull (1200 mL).

	Far-Field Focusing			**Near-Field Focusing**	
	Static	**Average**	**Filtered**	**Averaged**	**Filtered**
Photogrammetry	186%	222%	503%	112%	128%
Convex Hull	38%	56%	192%	2.5%	10%

overestimation in grape volume from 222% to 112% compared to the photogrammetry scan or from 56% to 2.5% compared to the convex hull of this scan. More work is needed to investigate how these results would vary with different volume estimation techniques from that used in this work or using a finer measurement grid spacing with the CNC.

It is worth noting that while it may be possible to determine true volume estimates using the acoustic techniques mentioned in this paper, the presented numerical volumes should only be considered as relative comparisons of the effect of each stage of the process. The establishment of an accurate relationship between acoustic volume and true cluster volume will require further study with a range of different grape clusters and foliage conditions. However, it should be noted that accurate measurement of the occluded grape volume using ultrasound is not necessarily essential. For example, it could potentially provide improved estimates of the proportion of occluded grapes to enhance yield estimates obtained using other methods such as computer vision techniques.

The process presented in this paper represents a significant improvement over the current state of the art ultrasonic methods for vine canopy assessment. The increased achievable detail will have a direct benefit for 3D volume estimation of vine canopies as well as improved ability to resolve potential grape clusters. These improvements should enable viticulturists to implement advanced precision viticulture techniques such as crop thinning, precise variable rate applications, and selective harvesting. We also anticipate that the techniques used will have applications beyond viticulture to other areas of horticulture.

7.1. Future Work

The lab results presented in this paper show promising initial results. However, field trials are needed to investigate how this system performs in a vineyard environment with different grapes varieties. The performance of the system needs to be investigated further with more leaves, grapes in closer proximity to the leaves, and occluding objects such as vine stems, trunks, and trellis materials. Solid obstacles such as trunks and trellis would not be disturbed by the agitation. This could be addressed by using a fusion of ultrasound and computer vision to assist in identifying these objects. Traditional computer vision techniques can be used to label visible areas of the ultrasonic scan such as vine stems, trunks, and other solid objects. This

could be extended further to develop an unsupervised machine learning process to directly classify regions of the acoustic recordings.

The effect of the presence of neighbouring grape clusters also needs to be investigated since they are likely to appear as a single larger cluster with the current processing and hardware. Scanning from different directions may provide improved ability to see behind solid objects or differentiate grape bunches which would otherwise be hidden by another cluster. Work is also needed to identify how early in the season this ultrasonic technique can be used to identify grapes bunch clusters and the relationship between the acoustic scan output and the true cluster weight.

Near-field focusing required additional processing overhead compared with far-field beamforming. One approach to address this may be to preclassify regions of the signal that contain significant reflected components and only perform beamforming on these regions. This could be assisted by incorporate a 3D depth cameras to provide additional information on where processing should be performed. Additionally, as each measurement location is independent of the others, simple parallelization techniques can be used to vastly improve processing times. In addition, improved resolution could be achieved by modifying the hardware so that the transmission could be focused in the near-field.

The array used in this study featured a very narrow beamwidth and could only image directly in front of the array. It used a highly accurate CNC machine to generate the 3D acoustic scans of the grapes. Although beyond the scope of this project, we believe we can enhance the hardware further to increase its practical use within a vineyard by reducing the scan time and remove the need for the CNC machine. Accurate tracking of the position of the array without the use of a CNC could be achieved using techniques such as a fusion of differential GPS and optical pose estimation.

If the array was redeveloped for large scale field trials, different transducers are likely to be used which may have different optimal transmitted signal. Therefore, it would be beneficial to further investigate the effect of transmitted waveform on the resulting scan and their resilience to sources of interference such as multipath reflections. Furthermore, given the significant physical differences between grape clusters and vine foliage, it may be possible to identify unique frequencies of absorption or reflection for each potentially making it possible to classify directly from the recorded waveforms.

Author contributions

Conceptualization, B.P and M.L.; methodology, B.P. and M.L.; software, B.P. and M.L.; validation, B.P. and M.L.; formal analysis, B.P. and M.L.; investigation, B.P.; resources, M.L and S.B; data curation, B.P.; writing—original draft preparation, B.P; writing—review and editing, M.L., F.A., and B.P.; supervision, M.L. and F.A.; project administration, M.L.; funding acquisition, M.L. and S.B

Funding

The researchers would like to acknowledge Bragato Research Institute (a subsidiary of New Zealand Winegrowers) as this research was supported in part by the Rod Bonfiglioli Scholarship.

Conflicts of interest

The authors declare no conflict of interest.

References

[1] Matese, A.; Gennaro, S.F.D. Technology in precision viticulture: A state of the art review. *Int. J. Wine Res.* **2015**, *7*, 69–81.

[2] Bramley, R.; Proffitt, A. Managing variability in viticultural production. *Grapegrow. Winemak.* **1999**, *427*, 11–16.

[3] Nuske, S.; Achar, S.; Bates, T.; Narasimhan, S.; Singh, S. Yield estimation in vineyards by visual grape detection. In *Proceedings of the Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference*; IEEE: New York , NY, USA, 2011; pp. 2352–2358.

[4] Nuske, S.; Gupta, K.; Narasimhan, S.; Singh, S. Modeling and calibrating visual yield estimates in vineyards. In *Field and Service Robotics*; Springer: Berlin/Heidelberg, Germany, 2014; pp.343–356.

[5] Mirbod, O.; Yoder, L.; Nuske, S. Automated measurement of berry size in images. *IFAC-PapersOnLine* **2016**, *49*, 79–84.

[6] Herrero-Huerta, M.; González-Aguilera, D.; Rodriguez-Gonzalvez, P.; Hernández-López, D. Vineyard yield estimation by automatic 3D bunch modelling in field conditions. *Comput. Electron. Agric.* **2015**, *110*, 17–26.

[7] Dey, D.; Mummert, L.; Sukthankar, R. Classification of plant structures from uncalibrated image sequences. In *Proceedings of the Applications of Computer Vision (WACV), 2012 IEEE Workshop*; IEEE: New York , NY, USA, 2012; pp. 329–336.

[8] Eccleston, K.W.; Platt, I.G.; Tan, A.E.-C. SAR for grape bunch detection in vineyards. In *Proceedings of the Microwave Symposium (AMS), 2018, Brisbane, Australia, 6–7 February 2018*; IEEE: New York , NY, USA, 2018; pp. 3–4.

[9] Gil, E.; Escola, A.; Rosell, J.; Planas, S.; Val, L. Variable rate application of plant protection products in vineyard using ultrasonic sensors. *Crop. Prot.* **2007**, *26*, 1287–1297.

[10] Llorens, J.; Gil, E.; Llop, J.; Escola, A. Variable rate dosing in precision viticulture: Use of electronic devices to improve application efficiency. *Crop. Prot.* **2010**, *29*, 239–248.

[11] Mazzetto, F.; Calcante, A.; Mena, A.; Vercesi, A. Integration of optical and analogue sensors for monitoring canopy health and vigour in precision viticulture. *Precis. Agric.* **2010**, *11*, 636–649.

[12] Tumbo, S.; Salyani, M.; Whitney, J.D.; Wheaton, T.; Miller, W. Investigation of laser and ultrasonic ranging sensors for measurements of citrus canopy volume. *APplied Eng. Agric.* **2002**, *18*, 367.

[13] Palacin, J.; Pallejà, T.; Tresanchez, M.; Sanz, R.; Llorens, J.; Ribes-Dasi, M.; Masip, J.; Arno, J.; Escola, A.; Rosell, J.R. Real-time tree-foliage surface estimation using a ground laser scanner. *IEEE Trans. Instrum. Meas.* **2007**, *56*, 1377–1383.

[14] Llorens, J.; Gil, E.; Llop, J. Ultrasonic and lidar sensors for electronic canopy characterization in vineyards: Advances to improve pesticide application methods. *Sensors* **2011**, *11*, 2177–2194.

[15] Palleja, T.; Landers, A.J. Real time canopy density estimation using ultrasonic envelope signals in the orchard and vineyard. *Comput. Electron. Agric.* **2015**, *115*, 108–117.

[16] Palleja, T.; Landers, A.J. Real time canopy density validation using ultrasonic envelope signals and point quadrat analysis. *Comput. Electron. Agric.* **2017**, *134*, 43–507.

[17] Kazys, R.J.; Vilpisauskas, A.; Sestoke, J. Application of air-coupled ultrasonic arrays for excitation of a slow antisymmetric lamb wave. *Sensors* **2018**, *18*, 2636.

[18] Allevato, G.; Hinrichs, J.; Rutsch, M.; Adler, J.; Jäger, A.; Pesavento, M.; Kupnik, M. Real-time 3D imaging using an air-coupled ultrasonic phased-array. In *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*; IEEE: New York, NY, USA, 2020; pp. 796–806. .

[19] Legg, M.; Bradley, S. Ultrasonic arrays for remote sensing of pasture biomass. *Remote. Sens.* **2019**, *12*, 111.

[20] Legg, M.; Bradley, S. Ultrasonic proximal sensing of pasture biomass. *Remote. Sens.* **2019**, *11*, 2459, doi:10.3390/rs11202459.

[21] Almqvist, M.; Holm, A.; Persson, H.W.; Lindström, K. Characterization of air-coupled ultrasound transducers in the frequency range 40 kHz-2 MHz using light diffraction tomography. *Ultrasonics* **2000**, *37*, 565–575.

[22] Parr, B.; Legg, M.; Alam, F.; Bradley, S. Acoustic identification of grape clusters occluded by foliage. In Proceedings of the Sensors and Applications Symposium (SAS 2020), Kuala Lumpur, Malaysia, 9–11 March 2020.

[23] DT9836 Series: High-Speed Simultaneous USB Devices with BNC. Available online: `https://www.mccdaq.com/Products/Multifunction-DAQ/DT9836` (accessed on 20 March 2021).

[24] Gan, T.H.; Hutchins, D.A.; Billson, D.R.; Schindel, D.W. The use of broadband acoustic transducers and pulse-compression techniques for air-coupled ultrasonic imaging. *Ultrasonics* **2001**, *39*, 181–194.

[25] ISO 9613-1:1993. *Acoustics—Attenuation of Sound During Propagation Outdoors—Part 1: Calculation of the Absorption of Sound by the Atmosphere*; International Organization for Standardization: Geneva, Switzerland, 1993; Available Online: `https://www.iso.org/standard/17426.html` (accessed on 20 March 2021).

[26] Yuan, W.; Zhou, T.; Jiajun, S.; Du, W.; Wei, B.; Wang, T. Correction method for magnitude and phase variations in acoustic arrays based on focused beamforming. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6058–6069.

[27] Camacho, J.; Martinez, O.; Parrilla, M.; Mateos, R.; Fritsch, C. A strict-time distributed architecture for digital beamforming of ultrasound signals. *IEEE Trans. Instrum. Meas.* **2010**, *59*, 2716–2723.

[28] Legg, M.; Bradley, S. Automatic 3D scanning surface generation for microphone array acoustic imaging. *Appl. Acoust.* **2014**, *76*, 230–237.

[29] Oxford Reference: Speed of Sound. Available online: `www.oxfordreference.com/view/10.1093/oi/authority.20110803100522606` (accessed on 20 March 2021).)

[30] Queiros, R.; Alegria, F.C.; Girao, P.S.; Serra, A.C. Cross-correlation and sine-fitting techniques for high-resolution ultrasonic ranging. *IEEE Trans. Instrum. Meas.***2010**, *59*, 3227–3236.

[31] Proakis, J.; Manolakis, D. *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd ed.; Prentice-Hall International Inc.: Upper Saddle River, NJ, USA, 1996.

[32] Ximin, Z.; Wanggen, W.; Li, X.; Junxing, M. Mean shift clustering segmentation and ransac simplification of color point cloud. In *Proceedings of the Audio, Language and Image Processing (ICALIP), 2014 International Conference*; IEEE: New York , NY, USA, 2014; pp. 837–841.

Chapter 6

Conclusion

Yield estimation is one of the key tools viticulturists use to inform management decisions, improve resilience to environmental conditions and optimise production. Yield estimation is traditionally a manual process relying on skilled staff to provide regular and timely insights. This is made more difficult by the spatial variability of vineyards and the subjectivity of individual assessments. Automating objective yield estimation within vineyards represents a crucial advancement in viticulture practice.

Modern computer vision approaches present an opportunity for solutions. However, they face difficulty with environmental constraints such as foliage occlusions, real-world lighting conditions, measuring the size and distribution of berries, and building solutions applicable to robotic automation. Researchers have proposed capturing 3D information as a means to addressing some of these issues.

Photogrammetry has been demonstrated as an effective method for capturing this information. However, the processing complexity means that real-time analysis is not possible. Alternatively, laser scanners have been used to reconstruct detailed phenotype representations of grape bunches. However, the expense and precise operation required their use in field conditions difficult and costly. In contrast, 3D cameras present a compelling option. They operate in real-time, which lends them to use in automated robotics applications, and their low cost makes them viable for commercial use. To date, there has been relatively little work towards their use for grape yield estimation in vineyards.

This study was the first to present a comprehensive comparative analysis of 3D camera technologies for precision viticulture. Specific attention was placed on performance characteristics that lend themselves towards accurately capturing the likeness of grapes and grape bunches.

In particular, we explored the ability of each technology to accurately represent the curvature of grapes surfaces, measurement errors against a ground truth 3D scan, and what proportion of a grape bunch can be reliably captured in different scenarios (fill rate). We showed that in most cases, structured light cameras, like the Kinect V1, struggled to capture the curvature of grapes surfaces, had poor fill rate even in ideal conditions, and had no ability to work in direct sunlight due to saturation. Active Infrared Stereoscopy cameras such as the Intel D435 worked significantly better. In indoor lab conditions, they presented the lowest average errors and best fill rates out of all technologies tested. However, it was noted that concave details between grapes were lost, reducing the contrast between individual grapes. Additionally, direct sunlight presented a challenge due to the saturation of the IR pattern. This required the camera to rely solely on the limited surface texture available and as a result performance dropped significantly.

In comparison, technologies that use time-of-flight principles such as the Kinect Azure DK and Intel L515 LiDAR, performed similarly regardless of ambient conditions. For both the ToF and LiDAR solutions tested, details between individual grapes were retained and there were no significant areas missing in scans; even in difficult direct sunlight conditions. However, we showed for the first time that the subsurface scattering of the emitted light causes distortions in the measured time of flight for each pixel. This manifests as peaks in the resulting 3D scan and in some cases causes grapes to be registered at greater depths than they actually are. Additionally, due to the different emission methods that ToF and LiDAR cameras use, distortions manifest slightly differently in each. ToF distortions are noticeably more peaky than observed in LiDAR.

These peaks present a distinguishable per-grape feature that was exploited in a novel approach for identifying grapes in ToF images captured by a consumer smartphone. It was found that while the distorted peaks were an adequate feature in most cases, they lacked prominence in grapes that were significantly occluded by other grapes in the bunch. Following this, we presented a novel approach that used automated peak detection, and 3D depth images to build a labelled dataset for training a state-of-the-art YOLO berry detection model in an unsupervised manner. The resulting model, while trained solely on images captured in the lab, performed well when applied to images captured of grape bunches in the field due to the unique approach of using 3D

thresholding to augment the backgrounds of lab-captured images. An R^2 value of 0.946 and an average precision of 0.970 was achieved when compared to manual counts. After detection, by exploiting the correlated depth images, matching peaks were able to be found for each identified grape. These 3D peak locations were used in a novel optimisation algorithm to estimate grape sizes and build a 3D representation of the visible portion of grape bunches.

We have shown that 3D cameras present an effective tool for solving some of the key challenges facing computer vision yield assessments. In particular, 3D cameras provide a significant advantage to unstructured berry size estimates. However, in situations where there are high levels of foliage occlusion, auxiliary approaches may be required. To this end, this work explored novel techniques for overcoming foliage occlusions using near-field beamformed ultrasound to construct detailed 3D volumes of vines. This is the first time that an ultrasonic phased array has been used to analyse canopy structures and the first time ultrasound has been used to visualise grape clusters. Through the agitation of foliage, we showed that in lab conditions it is possible to filter out leaves and identify volumes pertaining to grape bunches. This work is the first of its kind and opens up many opportunities for future research.

This work resulted in the publication of two peer-reviewed Q1 Journal Articles, with the third article currently in review, and a conference paper. A summary of the main contributions of this work is provided below:

1. A novel approach to analysing the performance of 3D cameras for the purpose of accurately capturing 3D curved surfaces.
2. The first study to benchmark the performance of multiple 3D camera technologies for use in grape yield estimation applications.
3. The first to show that diffused scattering of light within berries results in distortions in scans of grapes for ToF and LIDAR depth cameras. A novel technique is developed to exploit this to identify individual grape berries.
4. YOLO unsupervised training and detection of size without the need of reference object.
5. The first application of a smartphone's built-in 3D camera for grape yield estimation.

6. A novel modelling technique to estimate the size of berries in a bunch using the berry's estimated 3D location.
7. The first work to use an air-coupled ultrasonic phased array and coded waveforms for the purpose of analysing vine canopies.
8. The first study to investigate the use of ultrasound for imaging fruit occluded by leaves and removing echoes from leaves by agitation of foliage with a fan.
9. A novel technique for improving the resolution of array-based cross-correlation for near-field echoes.

The results of this study show that real-time 3D cameras have the potential to significantly contribute to the problems facing automated yield estimation for precision viticulture. The ability of 3D cameras to understand the depth and scale of objects improves the practicality of computer vision solutions in unstructured field conditions. We have shown that the current generation of 3D cameras is adequate at capturing information about individual grapes and that this can be exploited in a real-world context to generate 3D reconstructions of grape bunches. Furthermore, we have shown that 3D cameras in modern smartphones are capable of presenting an objective analysis of grapes bunches in the field without structuring the environment typically required to understand scale within 2D images. This drastically reduces the complexity of the data capture process and we expect it will have follow-through benefits to staffing and training requirements of viticulture staff. We have also demonstrated for the first time that air-coupled ultrasound arrays are effective at capturing the detailed structures of vines, and potentially differentiate between. Traditionally ultrasound has been limited to coarse estimates of foliage volume easily impacted by outliers. Thus, this new technique presents a significant improvement over state of the art.

By addressing the challenges associated with vineyard yield estimation, such as accuracy, object identification, occlusion, and manual sampling, this research contributes to improving vineyard management and profitability. It opens avenues for further advancements in computer vision, 3D camera technologies, and ultrasound volumetric scanning, paving the way for more efficient

and precise yield estimation in vineyards.

Future Works

This work presents significant advancements to the state of the art in understanding the capabilities of 3D cameras for grape yield estimation and use of ultrasound in viticulture. While the study treats ultrasound and 3D camera techniques separately, the possibility of combining them into a unified system should be considered once they reach similar Technology Readiness Levels (TRL). Potentially, such a system could be incorporated into an automated robotic platform; ultrasound to address occlusion and compute foliage volumes, and real-time 3D cameras to estimate accurate bunch and berry metrics. Towards this goal, there are exciting areas in both 3D imaging and ultrasound that warrant further research.

Yield Estimation with 3D Cameras

There is significant potential for ToF and LiDAR 3D cameras in real-world field applications due to their resilience and performance in direct sunlight. Furthermore, the unique distortions present in their scans of translucent grapes may be able to be exploited to perform nondestructive analyses of the properties of individual berries beyond the initial results presented in this work. More work is needed to evaluate if the unique distortion profiles of each berry correlate with physical characteristics such as berry size and orientation. Additionally, could the distortions be used to analyse the ripeness of grapes pre-harvest and to measure the degradation of the quality of grapes post-harvest? The potential value may be found by comparing differences in distortion profiles between different frequencies of transmitted light or the measurement techniques used. For example, in this work, we demonstrate that swept single-beam LiDAR and global illumination/shutter ToF systems exhibit different distortion profiles.

This work presents a novel approach to reconstructing 3D bunches and estimating berry sizes using solely the position of distorted peaks found in 3D scans. Empirical evaluation suggests that these estimates are close; however, more work is needed to compare them with physical

measurements. Additionally, this process can be extended with restricted reconstruction grammars to reconstruct the internal bunch structure and estimate the non-visible berries. Complete reconstructions of grape bunches will enable the determination of standard OIV descriptors such as cluster compactness, berry size variance, and cluster shape.

The real-time nature of the 3D cameras used in this work also lends itself well to multi-perspective reconstructions of the target bunch to increase the proportion able to be analysed. Simultaneous localisation and mapping (SLAM) techniques may be a suitable way forward in this respect. Such an approach could be extended to automating perspective changes with robotics, drawing inspiration from inherent human behaviour. When presented with a similar occlusion problem, a human's initial instinct is likely to change their perspective. Applied to robotics automation, it would improve understanding of individual bunches and provide a method for programmatically avoiding occlusions.

Current 3D cameras still feature various performance issues, from noise & flying pixels to multi-path distortions. In the case of stereo vision, lack of texture in objects surfaces, limited resolution, and changing ambient illumination can be detrimental to a successful outcome. Researchers have demonstrated various techniques to enhance the performance of low-cost 3D cameras such as shape from shading or shape from polarization, that may have a tangible impact on their suitability for grape yield estimation.

Ultrasonic Detection of Occluded Grapes

The novel ultrasonic techniques presented in this work represent the first time ultrasonic phased arrays have been used for precision viticulture. The unique near-field focusing we employ to reduce beamwidth as a post-processing step and the agitation-based variance filtering is also novel in its own right. The results obtained demonstrate promising potential but also indicate several avenues for future research.

Firstly, field trials are necessary to assess the performance of this system in real vineyard environments with varying grape varieties, denser foliage, and occluding objects such as vine stems,

trunks, and trellis materials. Integration with traditional computer vision techniques could enhance the analysis of volumetric scans, enabling the identification of non-leaf occlusions and the classification of ultrasonically scanned volumes.

Furthermore, 3D cameras could be used to facilitate performance improvements in near-field focusing by reducing the depth ranges needing to be processed during beamforming. Significant improvements could also be made by adopting new array designs, introducing transmission focusing, and parallelising computational and signal processing.

Support for large-scale data capture in vineyards may be approached by accurate tracking of the position of the array using techniques such as a fusion of differential GPS and optical pose estimation. Finally, given the distinguishable physical characteristics between grape clusters and vine foliage, it might be feasible to identify unique frequencies of absorption or reflection for each, potentially enabling direct classification from recorded waveforms.

The 3D volumetric scans enabled by ultrasonic phased arrays have potential in variable rate applications. Ultrasound is already commercially used as a method for real-time evaluation of foliage density. However, the transducers have a wide field of view and estimate foliage density by measuring the distance to the closest object. This makes them susceptible to overestimating the foliage present. The multi-echo volumetric scans presented in this work have the potential to provide a more accurate understanding of foliage and allow treatments to be applied more sparingly.

The use of ultrasonic phased arrays also has benefits beyond precision viticulture. For example, the array used in this work has been successfully demonstrated as a tool for pasture biomass estimation. We expect that similar techniques as presented in this work would be equally beneficial in other cropping systems, such as kiwifruit, tomatoes, and citrus crops.

Appendices

Appendix 1

This article was published and presented at the 2020 IEEE Sensors Applications Symposium (SAS) held remotely on March 9-11 from Kuala Lumpur, Malaysia. This work led to the extension presented in Chapter 5.

Acoustic Identification of Grape Clusters Occluded by Foliage

Baden Parr, Mathew Legg, Fakhrul Alam
Department of Mechanical and Electrical Engineering
School of Food and Advanced Technology (S&FAT)
Massey University
Auckland, New Zealand
{b.parr, m.legg, f.alam}@massey.ac.nz

Stuart Bradley
Inverse Acoustics Ltd
Auckland, New Zealand
inverse.acoustics@gmail.com

***Abstract*—The performance of a vineyard can be influenced by accurate yield estimations prior to harvest. Traditionally, this is a manual process. However, due to the high labour costs and subjective nature of manual assessments, researchers have been working on automated techniques. Utilising 2D computer vision has shown promising results but is inherently limited due to occlusions. The algorithms can only count grapes that are directly visible. Often this shortcoming is accounted for by using coarse occlusion ratio estimates, which themselves need to be manually determined. As a result, researchers have begun looking at alternative methods of grape detection. Synthetic Aperture Radar (SAR) has been demonstrated as a feasible approach to see grape clusters behind leaves. However, this comes at a significant financial cost. This paper introduces an alternative approach that utilises low frequency ultrasound to detect grape clusters in the presence of foliage occlusion. We demonstrate that such low frequency signals have the ability to propagate through foliage and reflect off grapes behind. Additionally, by agitating the leaves we can analyse the variance of consecutive samples and determine which volumes are likely to belong to grape clusters.**

***Index Terms*—ultrasound, vine yield, smart agriculture, non-destructive, remote-sensing**

I. Introduction

The profitability of vineyards can be influenced by the ability to estimate crop yield before harvest. With accurate yield estimations, viticulturists can implement precision agriculture techniques such as selective harvesting, crop thinning, or variable rate applications [1].

Traditionally, yield is estimated using manual techniques [2]. These approaches are labour intensive and prone to human error [3]. Additionally, obtaining enough samples to describe the variation within the vineyard can be difficult [4].

A 2001 report suggested that current sampling practices resulted in yield forecasts that were on average 33% from the actual [5]. The authors suggested that an improvement from 33% to 20% would result in an $85 million/year benefit to the Australian industry.

Computer vision has been explored as a potential solution for automating yield estimation. These techniques involve the capturing of high-resolution images of entire rows of vines; a process that can be manually assisted or entirely automated. Typically, images are taken from cameras mounted to the sides of farm machinery, such as all-terrain vehicles. This allows large sections of vine to be imaged with relative ease.

Numerous computer vision algorithms have been proposed to detect visible berries within each image. From these, an estimate of the yield and its variability can be determined. State-of-the-art algorithms have demonstrated detection of up to 92% of visible grapes. However, visible grapes are often only a subset of the total grapes on the vine. Furthermore, depending on the cultivar, training structure, and pruning techniques, the distribution of grape clusters is often not uniform throughout the vine. In addition, clusters that are visible are often occluded from other sources such as branches and leaves leading to misclassification and miscounts. Consequently, there is still a significant margin of uncertainty within the resulting estimations that cannot be easily solved using the current processes.

Individually, grapes do not reveal much about the performance of a vineyard. Of more interest are metrics describing the grape clusters. According to the OIV descriptor list[1], a grape clusters size, compactness, and shape are vital to understand vine growth and performance [6]. This can pose a problem when grape clusters are occluded from the imaging sensors field of view by some object such as a leaf or branch.

Traditionally, occlusions are accounted for by a simple scaling visibility factor, derived from manual measurements [7]. However, this approach is naive, and its accuracy will likely vary throughout a vineyard. Furthermore, a simple scaling factor only accounts for yield-based variations. Important metrics such as a clusters size and shape will remain affected. In some situations, images will be taken after manual removal of foliage around the grapes [8], [9]. However, this is not ideal for large scale autonomous solutions as it requires significant manual involvement. Occlusion may be the biggest unsolved issue when it comes to computer vision solutions for yield estimation.

The development of microwave-based yield estimators have been presented as a possible solution to the problem [10]. However, these technologies are costly and far from com-

[1]The OIV descriptor list summarizes established phenotypic traits of grapevines and grape clusters.

mercial implementation. Ultrasound has primarily found use in the agricultural industry for the measurement of crop canopies [11], [12]. These measurements can be used as coarse estimates of crop yield [13].

Within viticulture, the ultrasonic analysis of vine canopy has been used to efficiently modulate the real-time application of vine management products [14]. In addition, researchers have demonstrated that independent scans taken over the growing season can be an effective approach to monitoring vine vigour [15].

To date, existing uses of ultrasound for viticulture make use of single ultrasonic transducers. These typically feature a wide beamwidth between 20 degrees and 60 degrees which gives them a very low resolution view of a vine [14]. Furthermore, these sensors only report the time to the first echo making them incapable of sensing anything beyond an occluding object. When coupled with their inherent wide field of view, this can also result in erroneous, over estimations of canopy volume [15]. This is due to the sensor always picking up the closest object in its view instead of capturing the average depth of what is visible.

To improve the resolution of ultrasonic systems, arrays of ultrasonic transducers can be used along with techniques such as beamforming that reduce the beamwidth [16]. However, such processes have not been used before in precision viticulture.

In recent research, Legg and Bradley have developed a new ultrasonic array for measurement through pasture [17]. It is able to generate a very narrow beam width for both transmission and reception. The low frequencies employed allow reflections through the entire depth of the grass so an accurate biomass can be estimated [18].

A similar technique could potentially provide a solution to the detection of grape clusters on vines. Low frequency ultrasound will enable visibility through vine foliage to identify grape clusters that would be otherwise missed by a traditional optical system. This paper describes preliminary testing of this hypothesis and demonstrates that ultrasound has the ability to detect grape clusters through foliage.

To the best of the authors' knowledge this is the first work to explore the ability of ultrasound to detect fruit occluded by foliage.

II. System Overview

We believe that emitting a highly directional low frequency ultrasonic signal towards a grape vine and observing the reflected waveform at a number of locations would enable us to identify volumetric regions where grape clusters may be. We hypothesise that the higher density grapes will provide an effective surface for an ultrasound wave to be reflected from, while the relatively low density leaves occluding the grapes will allow the transmission of this wave. As the ultrasound will also reflect off the leaves, it then becomes a classification problem to identify which reflection most likely belongs to the grape clusters. This can be alleviated by slight agitation of the leaves by the light movement of air. The agitation will cause the reflected signal from the leaves to distort between consecutive samples while the relatively heavier grapes should remain stationary. Successive samples can then be analysed to statistically determine where grape clusters might be.

III. Experimental Hardware

Tests were conducted in lab environments utilising a highly directional narrow beamwidth ultrasonic array (see Fig. 1 & 2) and a computer controlled Cartesian gantry (CNC) that was mounted vertically. To perform relevant tests, clusters of table grapes were used in conjunction with a young grape vine with a substantial level of foliage. This setup can be seen in Figure 3.

A. Ultrasonic Array

The ultrasonic array used in this study was custom designed for precision agriculture requirements. It features an optimised array of 160 Murata MA40H1S-R surface mount ultrasonic transducers[2] and 204 MEMS microphones. The ultrasonic transducers are arranged in 9 rings each containing 17 transducers. Each ring is wired in parallel and can be operated independently to dynamically change the beamwidth of the device. A similar arrangement of 12 independent rings is used for the microphones. These transducers have a centre frequency of 40 kHz and a response around this that drops 20 dB either side at 25 kHz and 60 kHz. All this combines to produce a device featuring a combined transmitted/receive beam width of up to 6.8° and a dynamic range of up to 33 dB. The measured beam pattern of the array can be seen in Fig. 2. MATLAB was utilised for generating the transmission pulse and processing the recorded samples from Data Translations DT9836 data acquisition devices[3]. Signals were transmitted

[2]https://www.murata.com/products/productdetail?partno=MA40H1S-R
[3]https://www.mccdaq.com/Products/Multifunction-DAQ/DT9836

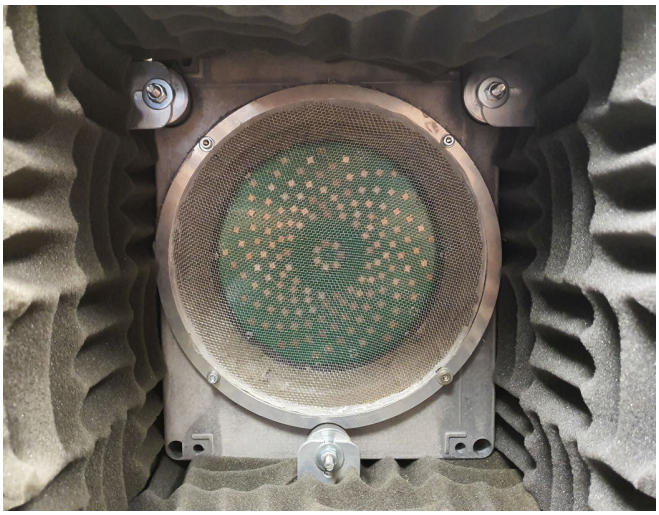


Fig. 1: View of the ultrasonic arrays main PCB installed in the measurement unit. The transducers can be seen in gold, arranged in an optimised spiral pattern.

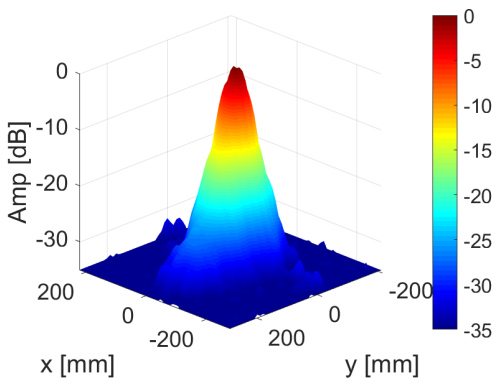


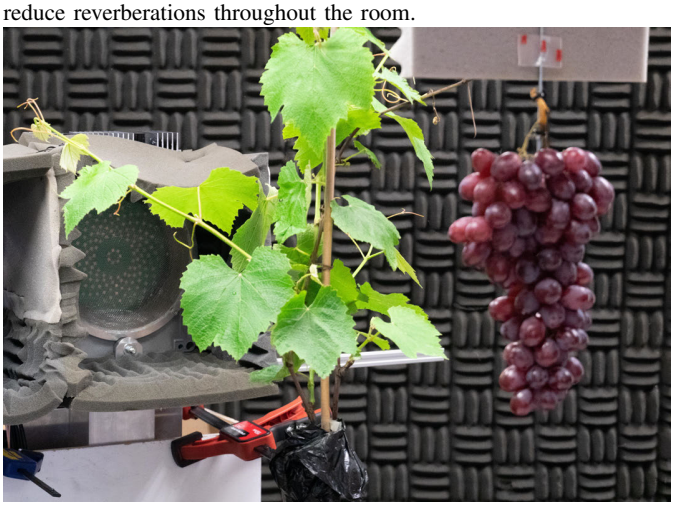
Fig. 2: Measured beam pattern for 35 kHz sine wave transmitted and received from the device.

at a 500 kHz sample rate and the 12 independent microphone channels were recorded at a sample rate of 225 kHz. The array was designed to be mounted to a farm vehicle. This made it too heavy to be directly mounted to the CNC machine. Instead both grapes and vine were supported from the CNC gantry and moved past the ultrasonic array. The exact details of the device are beyond the scope of this paper. For more information the reader is directed to the recent article by Legg and Bradley [19].

B. CNC Measurement Platform

The CNC used in this study has a accuracy of 0.025 mm and a usable range of motion of 1.4 m x 1.4 m. It is traditionally used in a more standard horizontal orientation. For this study, a custom gantry tower was built that could support the weight of the grapes and vine and positioning them up to 400 mm away from the face of the CNC machine. Grapes were supported 300 mm below the gantry tower using a 3 mm threaded steel rod. This was chosen to provide some level of resistance to motion when the CNC moves while also having a low profile from the perspective of the ultrasonic transducer. To support the vine from the gantry tower, its weight had to be reduced. To achieve this, it was removed from the planter and the majority of soil cleared from around its roots. The roots were then placed in a black plastic bag which was then tied around the base of the vine. The vine was supported with a bamboo pole which was cable-tied to the base of the plant and extended above the highest point of the vine where it was fixed to the gantry tower. This was important to ensure the vine itself was living and free-standing as in a field environment. The vine was positioned in such a way that there was a 200 mm gap between the closest leaf and the grape cluster. Acoustic foam was used to dampen reflections from the aluminium frame to either side of the gantry tower as well as in front of the metal supporting rod for the grapes. The ultrasonic transducer was positioned perpendicularly 1100 mm from the path of the grapes. Care was taken to ensure swinging movement caused by changing measurement position had halted before ultrasonic measurements were taken. Furthermore, the as the acoustics of the room itself was static to the array, care was taken to soften any strong reflective surfaces with noise dampening foam.

(a) A small vine was positioned in front of a cluster of grapes supported by a custom laser cut gantry tower fixed to a vertically orientated CNC machine. Acoustic damping foam was employed to reduce acoustic reflections from aluminium CNC components and reduce reverberations throughout the room.



(b) The positioning of the acoustic camera (background), small vine (mid-ground), and grape cluster (foreground). The grapes were hung using 3 mm threaded rod, and the vine by the bamboo stake it had grown around.

Fig. 3: The experimental set up used in this study. The images show how a cluster of grapes was positioned behind a small vine hung from a Cartesian CNC machine.

IV. Experimental Methodology

A. Ultrasonic Signal Processing

The optimal transmission waveform is a future research focus. However, a simple analysis was conducted to compare the performance of pulsed sine-waves, to linear chirped signals of varying duration and frequency. It was empirically determined

that a single 1.5 ms linear chirp from 35 kHz to 60 kHz was the best compromise among the frequency response of the transducers, the acquirable resolution of the grapes, and the penetrating performance of the signal. This signal was generated in MATLAB and a hamming window was applied before it was transmitted. Equation 1 shows how this signal was constructed.

$$s(t) = w_0^\tau(t) \cdot \sin\left(2\pi\left[f_0 + \frac{B}{2}\frac{t}{\tau}\right]\right) \cdot \Pi\left[\frac{(t-\tau/2)}{\tau}\right] \quad (1)$$

where t is time, f_0 is the frequency at $t = 0$, B is the bandwidth, τ is the pulse duration, $\Pi[(t-\tau/2)/\tau]$ is the rectangular function from 0 to τ, and $w_0^\tau(t)$ is a Hamming window applied to the width of the pulse.

The reflected signal was recorded by the 12 rings of microphones. These channels were averaged together to achieve the desired narrow beam width. Following averaging, the signal was then filtered with a 40 kHz bandstop filter in order to remove 40 kHz ringing at the resonance frequency of the transducers.

Cross-correlation was used to further improve the temporal resolution of response. This was conducted as shown in (2),

$$r_{hx}[n] = \sum_{k=-\infty}^{\infty} x[k]h[k-n] \quad (2)$$

where r_{hx} is the resulting cross correlation, x is the average of the 12 recorded signals, and h is a filtered version of the signal that was transmitted. Pre-filtering the transmitted signal was performed by a bandpass filter designed to emulate the frequency response of the transducers.

B. Measurement Procedure

The gantry tower was moved in a predetermined pattern covering an area 460 mm wide (x-axis) and 400 mm high (y-axis). This area was chosen to ensure full coverage of the grape cluster as it moved in front of the transducer. An x-axis step of 20 mm and y-axis step of 50 mm was used, resulting in a total of 216 measurement locations. The larger 50 mm y-axis step size was used to reduce number of samples required as we were primarily interested in the horizontal resolution due to the smaller cross section of the grape cluster from that perspective. When instructed by MATLAB, the CNC moves to a new position, pauses for 3 seconds and takes a sequence of 20 measurements. The entire 216 measurement run took approximately 20 minutes to complete. It should be noted that this time requirement is not a limitation of the technology, instead a limitation of the experiment platform. As we had to move the grapes and vine instead of the transducer we had to wait for any swinging movement to stop before measurements could commence.

Measurements were first taken without the vine and leaves present to establish an understanding of the grapes alone. Following this, the vine was added to the gantry and measurements were taken with and without agitation of the leaves. Agitation was performed with the use of a desk fan in a fixed position relative to the acoustic transducer. The fan was set to its lowest setting and directed towards an area directly in front of the acoustic transducer so to lightly disturb the vine foliage when it passed in front. In a field environment this agitation could be performed naturally by the wind, or supplemented by a similar portable fan.

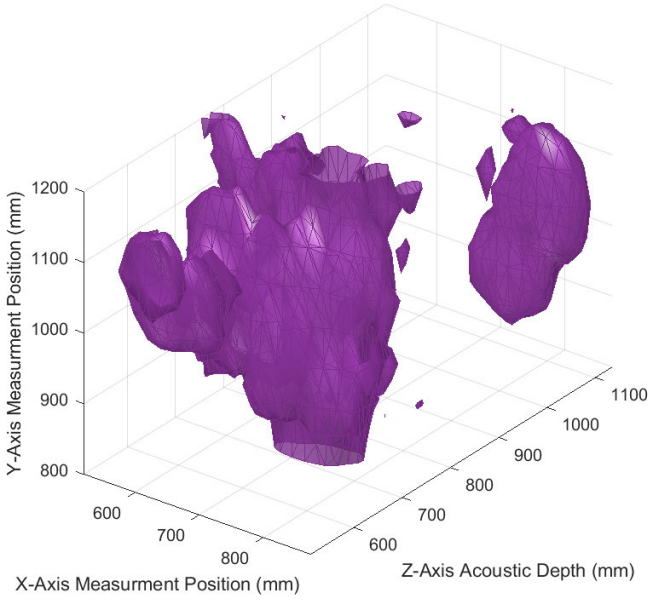


Fig. 4: An isosurface plot of a single sample RMS acoustic volume. The plot shows a significant volume of leaves in the foreground.

V. Results and Discussion

The resulting cross-correlated signals were mapped to an amplitude over distance representation using the speed of sound through air, $c = 343$ ms^{-1}. For each measurement location, a moving window was passed over the response which calculated the RMS of the signal for that position in space. The RMS for the nth window $R_w[n]$, over the cross-correlated signal r_{hx} is shown in (3),

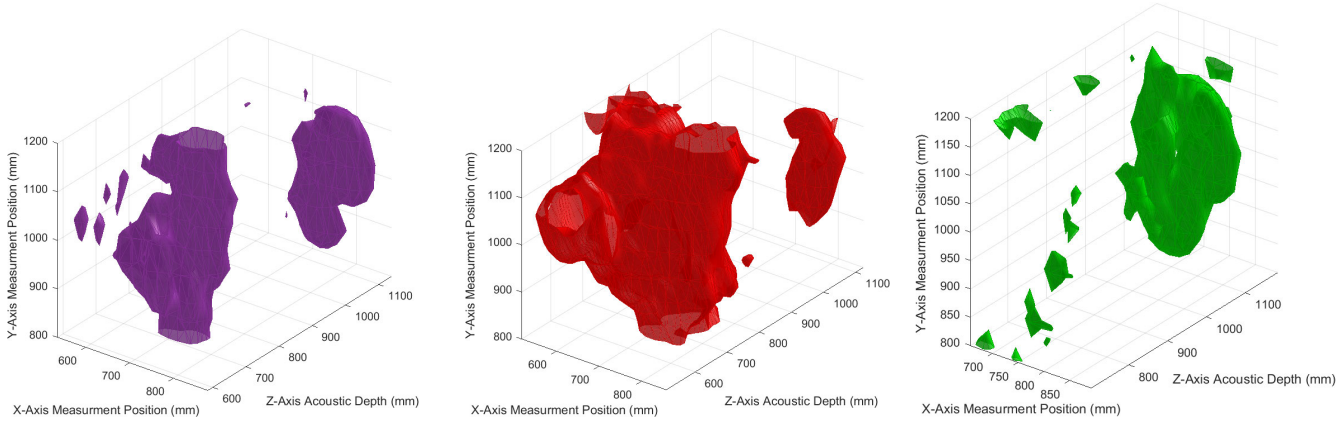
$$R_w[n] = \sqrt{\frac{\sum_{k=0}^{A} r_{hx}^2[B \cdot A + k]}{A}} \quad (3)$$

where A is the size in samples of the window used and B is the percentage of window overlap. In this study these were 26 and 0.5 respectively, corresponding to a depth window of 20 mm and an overlap of 10 mm.

The window was computed over the distance 500 mm to 1400 mm resulting in an RMS volumetric representation of the acoustic scan with the dimensions 900 mm×460 mm×400

TABLE I: Voxelised volume estimates of the foliage and grape components of each experimental scan.

Volume $(10^{-5}m^3)$	**Static Scan**	**Averaged Scan**	**Filtered Scan**
Leaf Volume	851	389	87
Grape Volume	166	187	350



(a) An isosurface of the RMS windows. Computed by averaging 20 recordings at each measurement point and computing the cross-correlation.

(b) An isosurface of the variation computed between 20 recordings at each measurement point within the volume.

(c) The result of binary filtering the averaged RMS volume with a thresholded variance volume.

Fig. 5: Isosurface plots showing the process used to establish grape clusters outlined in this study. The images show how a cluster of grapes, occluded behind a grape vine, can be effectively isolated and identified.

mm, represented by 89×24×9 uniformly spaced points. The lower limit of this volume was chosen due to a dead zone between 0 and 500 mm where the ultrasonic array is directly receiving the transmitted signal and subsequent PCB harmonics. The upper limit was chosen as beyond 1.4 m was free space and cropping this data out reduced subsequent processing required. Additionally, vineyard rows are typically around two meters wide which limits the expected range an ultrasonic array would need to operate at.

The computed volumes are visualised in this study by isosurfaces set to a threshold of 10% the maximum RMS of the entire scan. This was empirically determined to be a threshold that best demonstrates the response. In further research more advanced approaches can be taken such as mean-shift or k-means clustering [20]. The numerical volume of each isosurface can be naively determined by treating the grid points as voxel cuboids and counting those that are over this threshold within a given region.

The first experimental measurement consisted of a static scan where the leaves were not agitated. For the analysis of this, no averaging was performed, only a single recording per point was used. The volumetric RMS representation of this can be seen in Fig. 4. The results show two primary volumes. The smaller volume at a depth of 1100 mm belongs to the grape cluster and the significantly larger volume, between 500 mm and 900 mm, belongs to the occluding vine. An estimate of the size of each of these volumes is given in the second column of Table. I. By itself, without prior knowledge, this data provides no way to characterise which volume may belong to grapes or foliage.

The experiment was run again while agitating the area in front of the ultrasonic array. For this, 20 recordings at each point were used. The hypothesis being, that by averaging multiple samples of the agitated foliage, the echo from moving leaves would be cancelled and the relatively stationary grapes would remain. The average between all 20 recordings was calculated on a sample by sample basis using the process in (4),

$$\mu[n] = \frac{\sum_{i=1}^{20} x_i[n]}{20} \tag{4}$$

where $\mu[n]$ is the average for the nth sample across all 20 recordings and $x_i[n]$ is the individual sample for the ith recording. The average was computed before cross-correlation and RMS windowing was performed. The result is shown in Fig. 5a. It can be seen that agitation and averaging has a significant impact on the results. The size of the volume corresponding with the foliage has reduced while that of the grapes has grown. This is verified by the volume estimates given in Table. I, where the leaf volume has dropped from 851 to 389 $10^{-5}m^3$ and the grape volume has increased slightly from 166 to 187 $10^{-5}m^3$. Unfortunately, the problem of characterising each volume remains.

To further analyse this result, the variance between the 20 consecutive recordings was computed. This was done on a per sample basis over the entire recording in the same way the average was computed. The process is shown in (5),

$$\sigma^2[n] = \frac{\sum_{i=1}^{20} (x_i[n] - \mu[n])^2}{20} \tag{5}$$

where $\mu[n]$ is the previously computed average for each nth sample and $\sigma^2[n]$ is the variance for the nth sample across all 20 recordings. In the same way that the RMS volume was calculated using the cross-correlation, an RMS volume was created using this variance data. The result can be seen in Fig. 5b. The plot shows a significant amount of variation in the region of the leaves along with some where the grapes weren't completely stationary.

The variance volume was then applied as binary threshold to the original RMS volume computed from the cross-correlation of the averaged signals. The result of this can be seen in Fig. 5c. Almost all of the leaf volume has been removed and more of the grapes volume has become resolvable. This can be verified by the volume estimates as well. The volume associated with the foliage has dropped to 87 from 389 $10^{-5}m^3$ and the grape volume has increase from 187 to 350 $10^{-5}m^3$. Furthermore, the foliage volume is now fractured into multiple smaller clusters which can be filtered out in subsequent processing stages.

VI. Conclusion

This paper introduced an approach that utilises low frequency ultrasound to detect grape clusters in the presence of foliage occlusion. Results have confirmed that such low frequency signals have the ability to propagate through foliage and reflect off grapes behind. Additionally, by agitating the leaves we are able to analyse the variance of consecutive samples and determine which volumes are likely to belong to grape clusters. Future research will explore advanced signal processing methods such as focusing the array at particular depths to increase sensitivity, exploring alternative transmission waveforms, and conducting experiments within field environments.

Acknowledgments

This research was supported in part by the NZ Winegrowers Association through the Rod Bonfiglioli Scholarship.

References

[1] J. Llorens, E. Gil, J. Llop, and A. Escola, "Variable rate dosing in precision viticulture: Use of electronic devices to improve application efficiency," *Crop protection*, vol. 29, no. 3, pp. 239–248, 2010.

[2] R. Bramley and A. Proffitt, "Managing variability in viticultural production," *Grapegrower and Winemaker*, vol. 427, pp. 11–16, 1999.

[3] P. Jeandet, R. Bessis, and B. Gautheron, "The production of resveratrol (3, 5, 4'-trihydroxystilbene) by grape berries in different developmental stages," *American Journal of Enology and Viticulture*, vol. 42, no. 1, pp. 41–46, 1991.

[4] R. Bramley, "Vineyard sampling for more precise, targeted management," in *First Australian Geospatial Information and Agriculture Conference, Sydney, Australia*, 2001, pp. 17–19.

[5] P. R. Clingeleffer, S. Martin, G. Dunn, and M. Krstic, "Crop development, crop estimation and crop control to secure quality and production of major wine grape varieties: a national approach," 2001.

[6] O. I. de la Vigne et du Vin, "Oiv descriptor list for grape varieties and vitis species," 2007.

[7] O. Mirbod, L. Yoder, and S. Nuske, "Automated measurement of berry size in images," *IFAC-PapersOnLine*, vol. 49, no. 16, pp. 79–84, 2016.

[8] M. Herrero-Huerta, D. González-Aguilera, P. Rodriguez-Gonzalvez, and D. Hernández-López, "Vineyard yield estimation by automatic 3D bunch modelling in field conditions," *Computers and Electronics in Agriculture*, vol. 110, pp. 17–26, 2015.

[9] D. Dey, L. Mummert, and R. Sukthankar, "Classification of plant structures from uncalibrated image sequences," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, 2012, pp. 329–336.

[10] K. W. Eccleston, I. G. Platt, and A. E.-C. Tan, "SAR for grape bunch detection in vineyards," in *Microwave Symposium (AMS), 2018 Australian*. IEEE, 2018, pp. 3–4.

[11] S. Tumbo, M. Salyani, J. D. Whitney, T. Wheaton, and W. Miller, "Investigation of laser and ultrasonic ranging sensors for measurements of citrus canopy volume," *Applied Engineering in Agriculture*, vol. 18, no. 3, p. 367, 2002.

[12] A. Schumann and Q. Zaman, "Software development for real-time ultrasonic mapping of tree canopy size," *Computers and Electronics in Agriculture*, vol. 47, no. 1, pp. 25–40, 2005.

[13] Q. U. Zaman, A. W. Schumann, and H. K. Hostler, "Estimation of citrus fruit yield using ultrasonically-sensed tree size," 2006.

[14] E. Gil, A. Escola, J. Rosell, S. Planas, and L. Val, "Variable rate application of plant protection products in vineyard using ultrasonic sensors," *Crop Protection*, vol. 26, no. 8, pp. 1287–1297, 2007.

[15] F. Mazzetto, A. Calcante, A. Mena, and A. Vercesi, "Integration of optical and analogue sensors for monitoring canopy health and vigour in precision viticulture," *Precision Agriculture*, vol. 11, no. 6, pp. 636–649, 2010.

[16] D. Döbler, G. Heilmann, and R. Schröder, "Investigation of the depth of field in acoustic maps and its relation between focal distance and array design," in *Proceedings of Internoise*, 2008, pp. 654–660.

[17] S. Bradley and M. Legg, "Ultrasonic remote sensing for precision agriculture," in *Proceedings of the 23rd International Congress on Acoustics, integrating 4th EAA Euroregio 2019*, Aachen, Germany, 9-13 Sep. 2019, pp. 4999–5004.

[18] ——, "Precision ultrasonic sonar for pasture biomass," in *The 12th European Conference on Precision Agriculture*, Montpellier, France, 8-11 July 2019.

[19] M. Legg and S. Bradley, "Ultrasonic proximal sensing of pasture biomass," *Remote Sensing*, vol. 11, no. 20, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/20/2459

[20] Z. Ximin, W. Wanggen, X. Li, and M. Junxing, "Mean shift clustering segmentation and ransac simplification of color point cloud," in *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*. IEEE, 2014, pp. 837–841.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Baden Parr
Name and title of main supervisor:	Dr Mathew Legg
In which chapter is the manuscript/published work?	Chapter 3
What percentage of the manuscript/published work was contributed by the student?	80

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the experiments, performed data collection and analysis and produced the first draft of the manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Parr, B.; Legg, M.; Alam, F. Analysis of Depth Cameras for Proximal Sensing of Grapes. Sensors 2022, 22, 4179. https://doi.org/10.3390/s22114179

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Baden Parr	Main supervisor's signature:	Mathew Legg Digitally signed by Mathew Legg DN: cn=Mathew Legg, c=NZ, email=m.legg@massey.ac.nz Date: 2023.08.07 10:16:14 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Baden Parr
Name and title of main supervisor:	Dr Mathew Legg
In which chapter is the manuscript/published work?	Chapter 4
What percentage of the manuscript/published work was contributed by the student?	75

Describe the contribution that the student has made to the manuscript/published work:

The candidate wrote the custom smart phone application, conducted the experiments, performed data collection, designed the algorithms, performed the analysis and helped produced the first draft of the manuscript.

Please select one of the following three options:

(●) **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

B. Parr, M. Legg, F. Alam. Grape Yield Estimation with a Smartphone's Colour and Depth Cameras using Machine Learning and Computer Vision Techniques. Computers and Electronics in Agriculture (2023) 213, 108174. https://doi.org/10.1016/j.compag.2023.108174

( ) **The manuscript is currently under review for publication**

Please provide the name of the journal:

( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Baden Parr	Main supervisor's signature:	Mathew Legg Digitally signed by Mathew Legg DN: cn=Mathew Legg, c=NZ, email=m.legg@massey.ac.nz Date: 2023.08.07 10:16:58 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Baden Parr
Name and title of main supervisor:	Dr Mathew Legg
In which chapter is the manuscript/published work?	Chapter 5
What percentage of the manuscript/published work was contributed by the student?	80

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the experiments, performed data collection and analysis and produced the first draft of the manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Parr, B.; Legg, M.; Bradley, S.; Alam, F. Occluded Grape Cluster Detection and Vine Canopy Visualisation Using an Ultrasonic Phased Array. Sensors 2021, 21, 2182. https://doi.org/10.3390/s21062182

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Baden Parr	Main supervisor's signature:	Mathew Legg Digitally signed by Mathew Legg DN: cn=Mathew Legg, c=NZ, email=m.legg@massey.ac.nz Date: 2023.08.07 10:17:26 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Baden Parr
Name and title of main supervisor:	Dr Mathew Legg
In which chapter is the manuscript/published work?	Appendix 1
What percentage of the manuscript/published work was contributed by the student?	80

Describe the contribution that the student has made to the manuscript/published work:
The candidate conducted the experiments, performed data collection and analysis and produced the first draft of the manuscript.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:
  B. Parr, M. Legg, F. Alam and S. Bradley, "Acoustic Identification of Grape Clusters Occluded by Foliage," 2020 IEEE Sensors Applications Symposium (SAS), Kuala Lumpur, Malaysia

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Baden Parr	Main supervisor's signature:	Mathew Legg Digitally signed by Mathew Legg DN: cn=Mathew Legg, c=NZ, email=m.legg@massey.ac.nz Date: 2023.08.07 10:15:36 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.