Automatic grape bunch detection in vineyards based on affordable 3D phenotyping using a consumer webcam

Thiago Teixeira Santos[1], *Luís Henrique Bassoi*[2], *Henrique Oldoni*[3], *Roberto Luvisutto Martins*[3]

[1]Embrapa Agricultural Informatics
Campinas, São Paulo, Brazil
thiago.santos@embrapa.br

[2]Embrapa Instrumentation
São Carlos, São Paulo, Brazil
luis.bassoi@embrapa.br

[3]Faculty of Agronomic Sciences, Universidade Estadual Paulista – Unesp
Botucatu, São Paulo, Brazil
heriqueoldoni@gmail.com, rlmartins.agro@hotmail.com

RESUMO

O presente trabalho apresenta uma metodologia para a análise fenotípica de vinhas por reconstrução 3-D, a partir de imagens obtidas por uma webcam de alta definição e baixo custo. Um novo software aplicativo integrou componentes de odometria visual e visão estéreo por múltiplas imagens para criar nuvens de pontos tridimensionais densas e precisas para as videiras, em escala milimétrica. Características geométricas e colorimétricas dos pontos foram empregadas em um procedimento de classificação que atingiu 93% de acurácia na detecção de pontos pertencentes às uvas. Os cachos individuais foram automaticamente delimitados e seus volumes estimados. A soma dos volumes estimados por videira apresentou um coeficiente de correlação de $R = 0,99$ ao peso real das uvas observado em cada videira após a colheita.

PALAVRAS-CHAVE: Viticultura, Estimativa de podução, Métodos não-invasivos, Fenotipagem 3-D, Visão estéro múltipla, SLAM.

ABSTRACT

This work presents a methodology for 3-D phenotyping of vineyards based on images captured by a low cost high-definition webcamera. A novel software application integrated visual odometry and multiple-view stereo components to create dense and accurate three-dimensional points clouds for vines, properly transformed to millimeter scale. Geometrical and color features of the points were employed by a classification procedure that reached 93% of accuracy

on detecting points belonging to grapes. Individual bunches were automatically delimited and their volumes estimated. The sum of the estimated volumes per vine presented a coefficient of correlation of $R = 0.99$ to the real grape weight observed in each vine after harvesting.
KEYWORDS: Viticulture, Yield estimation, Non-invasive methods, 3-D phenotyping, Multiple view stereo, SLAM.

INTRODUCTION

Viticulture is an agricultural activity composed by high value perenial crops. Precision viticulture, yield prediction and vine breeding are examples of demanded applications that need some sort of in field phenotyping, but find a bottleneck in laborious measurement by human operators or destructive and sparse sampling. Non-invasive computer vision-based systems have been proposed in research literature, raising as a promising alternative for vineyard phenotyping, and exploiting the ubiquity of digital cameras and their capacity to acquire large amounts of data.

Two different approaches are found in the literature. The 2-D approaches perform grape detection directly on single images, as seen in the works of Nuske et al. (2014). Such methods benefit of their simplified image acquisition step, but occlusions caused by leaves, branches and other bunches are a drawback. An alternative approach are the 3-D based techniques that employ multiple-view stereo vision (HARTLEY; ZISSERMAN, 2004) to create three dimensional models of grapevines in field. Such three-dimensional methods present two clear advantages: they can handle occlusions better than 2-D alternatives, at the cost of a more complex image acquisition step, and they can provide volumetric information, useful on estimating biological attributes as mass, form and fruit quality.

Herrero-Huerta et al. (2015) used a photogrametric software developed by themselves for multiple-view stereo vision, producing sets of 3-D color points, *point clouds*, for vines in field. Convex hulls and solid models were then employed to estimate the bunches volumes and weights. The authors reported a coefficient of determination around 0.77 between their estimated values and the measured biological traits (weight, volume and number of berries). However, their methodology apparently lack an automatic grape detection procedure. Rose et al. (2016) employed a robot carrying a five camera system and a lighting unit to take images of vineyards line at night. The authors used a commercial multiple-view stereo software to produce point clouds. Grape detection was performed using a kernel-based classifier, geometrical features, the *Fast Point Feature Histograms* (FPFH) proposed by Rusu, Blodow e Beetz (2009), and color features in the HSV colorspace. Authors report around 80% of recall and precision for bunches classification in Riesling vineyards cultivated under a vertical trellis system. Another interesting approach is the method proposed by Klodt et al. (2015), an intermediary solution between the 2-D and 3-D approaches: pairs of images are employed to create dense depth maps used in grape classification.

The method proposed in this work is a 3-D based approach. Point clouds were produced

using a novel software, developed by the authors, that combines visual odometry and multiple-view stereo components, resulting in dense, accurate and metric 3-D models for vines in field, from images captured by a simple consumer grade HD webcam. Similar to Rose et al. (2016), FPFH and color features were combined and a support vector classifier was used to detect points belonging to the grapes surfaces. Finally, the volume of grapes was computed for each vine and a regression performed between the estimated volume and the real weight after harvesting.

MATERIALS AND METHODS

Plant material

Five vines of Chardonnay were selected in the vineyards of Guaspari Winery, located at Espírito Santo do Pinhal, São Paulo, Brazil (Lat -22.181018, Lon -46.741618). The winery staff performed dual pruning: one for shaping (after previous year harvest) and one for production, resulting in canopies of lower density. The data gathering was performed few hours before the harvest in May 16, 2017. Images were captured, then bunches were collected and weighted in a scale.

Image data acquisition and 3-D reconstruction

The image data acquisition system was an affordable solution developed by the authors to meet the the idea *affordable and lean phenotyping* for plant science (FIORANI; SCHURR, 2013). The hardware part consisted in a quadri-core computer (an Intel® Core™ i7-based consumer grade notebook) and a HD webcam (a Logitech® c920), able to capture 1920×1080 pixels color frames. The software part consisted in a GNU Linux-based system including:

- a modified version of the ORB-SLAM visual odometry system proposed by Mur-Artal, Montiel e Tardós (2015);
- the patch-based multiple-view stereo system (PMVS) proposed by Furukawa e Ponce (2010), and
- an application implementing the proposed methodology for 3-D reconstruction, *3-Demeter Capture*.

3-Demeter Capture implements an improved version of the methodology previously proposed by Santos e Rodrigues (2016). The camera was set for fixed focus and calibrated using the Zhang (2000) calibration method. This calibration provides the camera *internal parameters* matrix $\mathtt{K}$ (HARTLEY; ZISSERMAN, 2004), immutable along the image acquisition process due to the fixed focus setting. To perform multiple-view stereo, the camera *external parameters*, ie, its rotation and translation parameters for each captured image are also needed. In the robotics and computer vision literature, such estimation of camera external parameters and the needed environment landmarks is called *simultaneous localization and mapping* (SLAM[1])

[1]Actually, the procedure employed in this work is a special case named *monocular visual SLAM* because it employs a single camera and no other sensors.

(SCARAMUZZA; FRAUNDORFER, 2011).

In Santos e Rodrigues (2016), the external parameters were provided by the SVO system proposed by Forster, Pizzoli e Scaramuzza (2014). SVO was able to perform the camera tracking needed for successful 3-D reconstructions, but presented some issues in those experiments: (i) camera tracking showed low robustness to fast camera rotations; (ii) posters containing textures were needed to provide the landmarks employed in camera pose estimation[2], and (iii) SVO implementation did not provide a *re-localization* procedure, able to reset the SLAM system to a sane state after camera tracking is lost, ie, after a pose estimation failure. Instead SVO, 3-Demeter Capture employs a modified version of the ORB-SLAM system (MUR-ARTAL; MONTIEL; TARDóS, 2015). ORB-SLAM has shown superior performance in the authors current phenotyping experiments, being able to find landmarks for SLAM in the surrounding environment, presenting better robustness for rotations and translations and a reliable re-localization module. In the present work, ORB-SLAM code was modified to recover the system *keyframes*. According to Mur-Artal, Montiel e Tardós (2015), ORB-SLAM selects from the video stream a set of frames (keyframes) that lead to robust SLAM, generate a compact and trackable map for the environment and avoid redundancy. This feature is exploited by 3-Demeter Capture for automatic selection of video frames, leading to accurate 3-D reconstructions in the following stereo vision step.

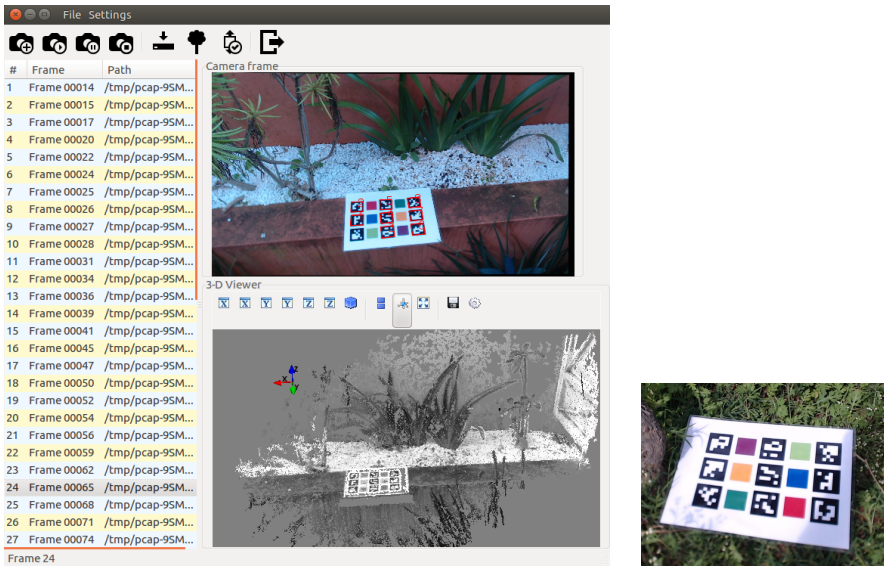
At the end of image acquisition step, 3-Demeter Capture presents a set of data composed by (i) the internal camera parameters matrix $\mathtt{K}$, (ii) a set of N images corresponding to video frames of 1920×1080 pixels and (iii) the camera pose for each image, in the form of rotation matrices $\{\mathtt{R_i}\}_{i=1..N}$ and translation vectors $\{\mathtt{t_i}\}_{i=1..N}$ (HARTLEY; ZISSERMAN, 2004; SCARAMUZZA; FRAUNDORFER, 2011). The application is able to export this data to PMVS and start the multiple-view stereo (FURUKAWA; PONCE, 2010; SANTOS; RODRIGUES, 2016). The point clouds produced by PMVS are actually a set of *surfels*. A surfel is a multidimensional data entry composed by:

- a 3-D point coordinate (x, y, z);
- a vector (n_x, n_y, n_z), corresponding to the normal of the object surface at the point, and
- a color triplet (r, g, b) in the RGB colorspace, corresponding to the observed reflectance of the surface at the point.

However, the *scale* of the produced point cloud is arbitrary. To address this issue, 3-Demeter Capture is able to perform an optional *normalization* step, composed by automatic scaling and rotation to a reference orientation (the Z axis pointing upwards). The application is able to look for a board containing easily recognizable markers presenting known geometry and size (Figure 1). The board geometry is used to transform the point cloud to millimeter scale, a procedure similar to the one employed by Herrero-Huerta et al. (2015). Also, the cloud is transformed to a standard reference frame, where the X and Y axes define the board's plane,

[2]The posters were needed in situations where the environment did not contained very textured objects in the camera field of view.

Figura 1: Screenshot of the 3-Demeter Capture application and the reference board. Application: (a - top) a camera frame where the board was found (the identified markers are highlighted in red); (a - bottom) the normalized 3-D point cloud (note the axis indicator). (b) The reference board with markers.



(a) (b)

and the Z axis is normal to the board. If the board is laid on the ground, the Z axis properly points upwards (see Figure 1).

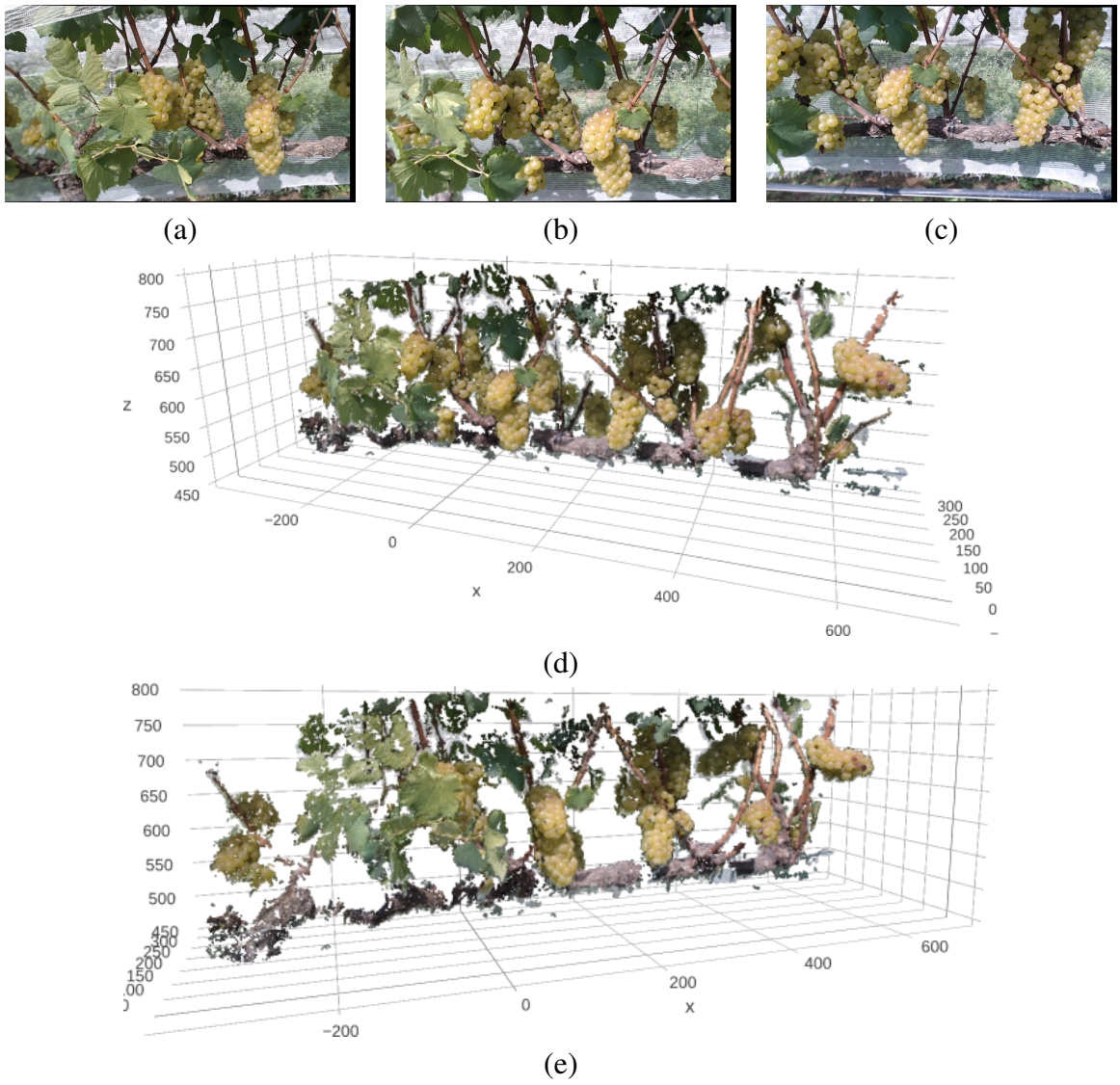
Points classification

The automatic detection of grapes bunches was designed as a supervised classification problem over the set of points in the point cloud. Similar to Rose et al. (2016), we combined geometrical and color information to characterize each point of the cloud as a feature vector. The geometrical features used were the FPFH proposed by Rusu, Blodow e Beetz (2009) and implemented in the Point Cloud Library (PCL) (RUSU; COUSINS, 2011). FPFH represents the geometrical properties of a point's neighborhood by generalizing the mean curvature around the point as a multi-dimensional histogram[3]. This descriptor has been used in several works on 3-D phenotyping for plant parts classification (PAULUS et al., 2013; WAHABZADA et al., 2015; ROSE et al., 2016). A FPFH feature vector is composed by a set of 33 values that are a characterization of the surface around a point: grape bunches, vine leaves, trunks and branches should present different patterns for this descriptor, according to the properties of their surfaces.

Color is also an useful feature for bunches classification. Klodt et al. (2015) used the RGB color space while Rose et al. (2016) employed the three components of the HSV colorspace. In this work, the a^* and b^* components from the CIELAB colorspace were employed. The color information is found in the a^* and b^* components while luminance is expressed on L^*. The L^* channel was discarded in an attempt to make the descriptor more robust to light variations.

[3]See Rusu e Cousins (2011) for a detailed description

Figura 2: Point cloud for a vine. Three of the frames collected by 3-Demeter Capture for this vine, (a), (b) and (c). The point cloud produced by PMVS and normalized by 3-Demeter Capture, seen in two different poses, (d) and (e).



A dataset composed by 121,758 points and their descriptors was employed for classification, under a supervised machine learning framework. The points were manually classified in three classes: (i) grape bunches, (ii) leaves and (iii) trunks and branches, composed by 58,421, 37,258 and 26,079 samples respectively. A support vector machine (SVM) classifier was employed for classification. The classifier penalty parameter C and the used kernel were selected automatically from a set of options by K-Fold cross-validation ($K = 3$) (HASTIE; TIBSHIRANI; FRIEDMAN, 2001). K-Fold cross-validation is employed again to measure the prediction performance of the classifier (using $K = 3$ again), in a *nested cross-validation* approach. The *scikit-learn* (PEDREGOSA et al., 2011) machine learning library was employed for SVM and cross-validation algorithms.

Bunches segmentation, filtering and volume estimation

Points classified as grape surface points were grouped in bunches by finding connected components. Two points were considered connected if the distance between then was up to 3 mm. A connected component is a set of points $\mathcal{B}$ such as for any two points $u_i, u_j \in \mathcal{B}$, there is a path $u_i \rightarrow v_0 \rightarrow v_1 \rightarrow ... \rightarrow v_{P-1} \rightarrow v_P \rightarrow u_j$ linking u_i and u_j, where $v_0, v_1, \ldots, v_P \in \mathcal{B}$ and any two adjacent points v_k, v_{k+1} are connected, ie, are up to 3 mm appart[4]. Each connected components was considered a *bunch*. To filter out noise caused by errors in classification, bunches containing less then 200 points were discarded.

The convex hull for the set of points $u_i \in \mathcal{B}_k$ is computed for each bunch $\mathcal{B}_k$. The convex hull volume is considered here an approximation for the volume of the real grapes bunch in the vine. The convex hull implementation available in SciPy (JONES et al., 2001–) was employed for hull and volume computation. Finally, the individual bunches volumes were summed up to produce a estimation for the vine total volume of grapes.

RESULTS AND DISCUSSION

The data acquisition using 3-Demeter Capture took 1 to 2 min per vine, producing around 90 images per vine (Table 1). PMVS took around 20 minutes per vine to produce a dense point cloud by multiple-view stereo. An example of 3-D reconstruction can be seen in Figure 2. Exploiting the fact the point clouds are properly oriented in a reference frame and in millimeter scale, points belonging to the ground or outside the region of interest were discarded (the grapes are found around 50 cm above the ground).

Tabela 1: Summary of data per vine.

Plant	Num. of images	Num. of 3-D points†	Estimated volume (cm^3)	Weight (g)
Vine 1	89	125,828	1,240.53	1,405.9
Vine 2	96	160,714	566.55	851.4
Vine 3	86	177,218	567.93	750.5
Vine 4	104	172,768	2,311.44	2,420.8
Vine 5	88	121,844	1,585.70	1,771.0

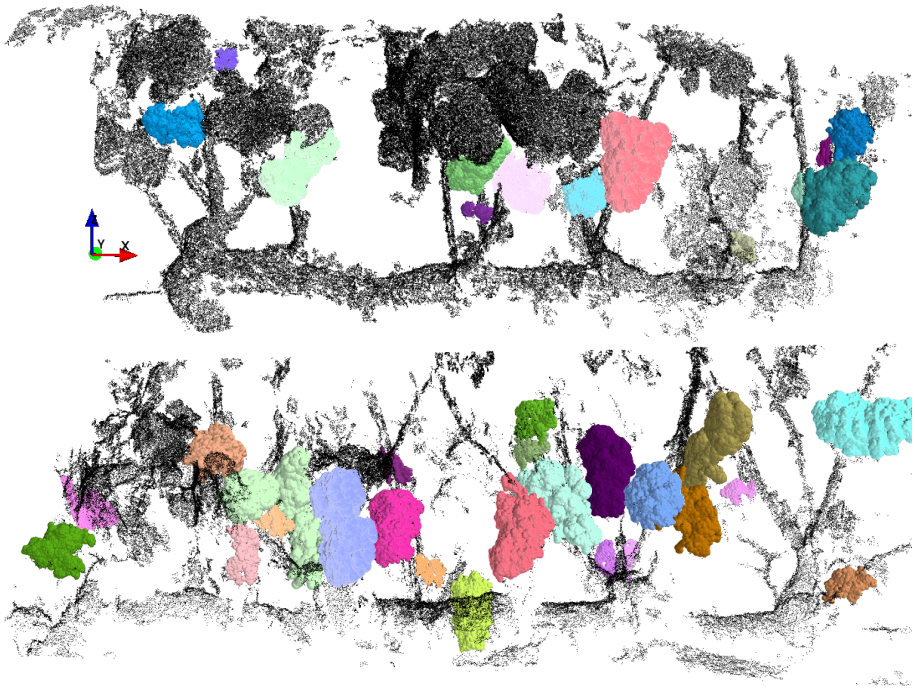
† After the selection of the region of interest.

Classification accuracy was estimated using 3-fold cross validation. The three estimations for accuracy were 93.02%, 92.96% and 93.16% on the annotated dataset, when using the SVM classifier and the 35-features descriptor (33 FPFH features plus the a^* and b^* values). It is interesting to note that using the SVM classifier over just the color descriptors (a^* and b^*) resulted in accuracy around 90% while the 33 FPFH features alone reached less then 70% of accuracy.

[4]This 3 mm threshold produced good results, considering the density of the points clouds.

Figure 3 shows the bunches found for vines 2 and 4. Each bunch is displayed in a different color while points corresponding to leaves and branches are shown in black. Figure 4 shows the linear regression between the estimated total volume of bunches and the true biological yield, the mass of grapes, measured using a digital scale after harvesting (see also Table 1). The correlation coefficient is 0.9984. A larger set of vines should be employed in future experiments to confirm such a good correlation, but this result is not unexpected: a volume estimation made using a 3-D model presenting millimeter precision, combined to a compact grape variety as Chardonnay, should be strongly correlated to weight.

Figura 3: Bunches automatically found in vine 2 (top) and vine 4 (bottom). Each bunch is shown in a different color. Points classified as non belonging to grapes are marked in black.

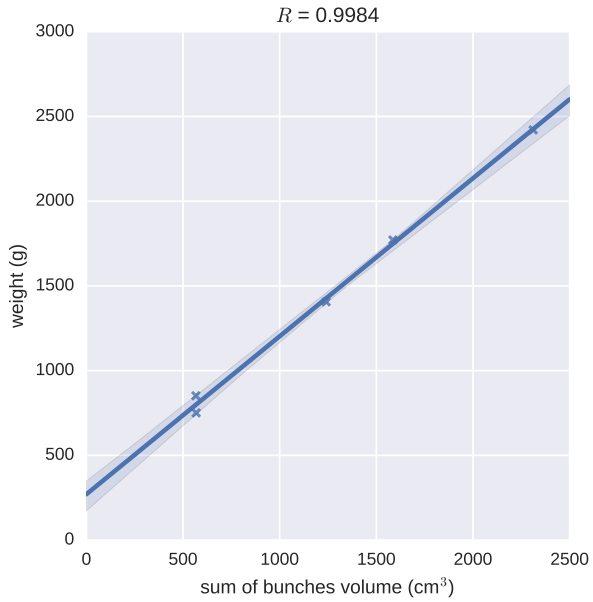


CONCLUSIONS

This work presented an affordable system able to (i) produce three dimensional models of grapevines in field; (ii) recognize bunches, leaves and branches structures in the plants, and (iii) detect bunches of grapes and compute their volume. The computed volumes presented excellent correlation to the vines yields. Applications on plant phenotyping, precision viticulture and agricultural robotics could benefit from these results.

A major advantage of non-invasive methods is the ability to perform non-destructive estimations, enabling the acquisition of developmental data over time. A possible future work would be the monitoring of the entire development cycle until the harvesting.

Figura 4: Regression between the vines estimated volumes and total mass of grapes.



ACKNOWLEDGMENTS

This work was supported by Brazilian Agricultural Research Corporation (Embrapa) under grant 01.14.09.001.05.04. We would like to thank Guaspari Winery for providing access to its wineyards and logistical support.

REFERÊNCIAS

FIORANI, F.; SCHURR, U. Future scenarios for plant phenotyping. *Annual review of plant biology*, v. 64, p. 267–91, jan 2013. ISSN 1545-2123.

FORSTER, C.; PIZZOLI, M.; SCARAMUZZA, D. SVO: Fast Semi-Direct Monocular Visual Odometry. In: *IEEE International Conference on Robotics and Automation (ICRA)*. [S.l.: s.n.], 2014.

FURUKAWA, Y.; PONCE, J. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 8, p. 1362–1376, Aug 2010. ISSN 0162-8828.

HARTLEY, R.; ZISSERMAN, A. Book. *Multiple View Geometry in Computer Vision*. 2. ed. [S.l.]: Cambridge University Press, 2004. ISBN 0521540518.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001. (Springer Series in Statistics).

HERRERO-HUERTA, M. et al. Vineyard yield estimation by automatic 3d bunch modelling in field conditions. *Computers and Electronics in Agriculture*, v. 110, p. 17 – 26, 2015. ISSN 0168-1699.

JONES, E. et al. *SciPy: Open source scientific tools for Python*. 2001–. [Online]. Disponível em: <http://www.scipy.org/>.

KLODT, M. et al. Field phenotyping of grapevine growth using dense stereo reconstruction. *BMC Bioinformatics*, v. 16, n. 1, p. 143, 2015. ISSN 1471-2105.

MUR-ARTAL, R.; MONTIEL, J. M. M.; TARDóS, J. D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, v. 31, n. 5, p. 1147–1163, Oct 2015. ISSN 1552-3098.

NUSKE, S. et al. Automated visual yield estimation in vineyards. *Journal of Field Robotics*, v. 31, n. 5, p. 837–860, sep 2014. ISSN 15564967. Disponível em: <http://doi.wiley.com/10.1002/rob.21541>.

PAULUS, S. et al. Surface feature based classification of plant organs from 3D laserscanned point clouds for plant phenotyping. *BMC Bioinformatics*, v. 14, n. 1, p. 238, 2013. ISSN 1471-2105.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

ROSE, J. et al. Towards Automated Large-Scale 3D Phenotyping of Vineyards under Field Conditions. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 16, n. 12, p. 2136, dec 2016. ISSN 1424-8220. Disponível em: <http://www.mdpi.com/1424-8220/16/12/2136>.

RUSU, R. B.; BLODOW, N.; BEETZ, M. Fast point feature histograms (FPFH) for 3D registration. In: *2009 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2009. p. 3212–3217. ISSN 1050-4729.

RUSU, R. B.; COUSINS, S. 3D is here: Point Cloud Library (PCL). In: *2011 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2011. p. 1–4. ISSN 1050-4729.

SANTOS, T. T.; RODRIGUES, G. C. Flexible three-dimensional modeling of plants using low- resolution cameras and visual odometry. *Machine Vision and Applications*, v. 27, n. 5, p. 695–707, 2016. ISSN 1432-1769.

SCARAMUZZA, D.; FRAUNDORFER, F. Visual Odometry [Tutorial]. *IEEE Robotics & Automation Magazine*, v. 18, n. 4, p. 80–92, dez. 2011. ISSN 1070-9932.

WAHABZADA, M. et al. Automated interpretation of 3D laserscanned point clouds for plant organ segmentation. *BMC Bioinformatics*, v. 16, n. 1, p. 248, 2015. ISSN 1471-2105.

ZHANG, Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 11, p. 1330–1334, Nov 2000. ISSN 0162-8828.