



COMPREENSÃO DE CENAS EM AGRICULTURA POR REDES NEURAI PROFUNDAS

Marcos Gabriel Barboza Duré **Diaz**¹; Thiago Teixeira **Santos**²

Nº 20601

RESUMO – A compreensão de cenas tridimensionais na agricultura é de interesse estratégico para a atividade, pois abre caminho para automatizar análises e processos produtivos no campo. Na última década, algoritmos de aprendizado profundo se tornaram estado da arte em tarefas de detecção e classificação em imagens. Ao mesmo tempo, algoritmos de reconstrução tridimensionais de estruturas a partir de imagens se tornaram cada vez mais robustos e escaláveis. Redes neurais profundas, no entanto, ainda são pouco aplicadas a dados tridimensionais, em especial a representações de cenas na agricultura. Neste trabalho, mostramos os resultados de testes de reconstrução tridimensional de linhas de uma vinícola e desenvolvemos a base de uma interface para anotação de nuvens de pontos, fundamental para o treinamento de redes neurais. Acreditamos que as redes profundas serão capazes de segmentar, classificar e identificar a qual instância pertencem objetos de interesse em videiras, como frutos, folhas e ramos. Esse resultado é aplicável no auxílio de atividades que precisem de análise fenotípica em campo, como robótica de precisão, acompanhamento da produção e previsão de rendimentos de cultivos.

Palavras-chaves: Redes Neurais Profundas; Reconstrução Tridimensional; Anotação de Dados; Classificação e Segmentação; Nuvens de Pontos

1 Autor, Bolsista FAPESP: Graduação em Engenharia de Computação, UNICAMP, Campinas-SP; marcos.dure.diaz@gmail.com

2 Orientador: Pesquisador da Embrapa Informática Agropecuária, Campinas-SP; thiago.santos@embrapa.br



ABSTRACT – *3D understanding of agricultural scenes is of strategic importance to productive processes. In the last decade, deep learning algorithms became the state of the art in image detection and classification. Concurrently, algorithms for reconstruction of 3D structure from images became more robust and scalable. However, deep neural networks have been seldom applied to 3D data, specially agricultural scene representations. In this work we show the results from testing 3D reconstruction on vineyards and develop the basis of a point cloud annotation interface, an important step to train neural networks. We believe deep networks will be capable of segmenting, classifying and identifying the instance of objects of interest in vineyards, such as fruits, leaves and branches. This result can be applied in activities that need phenotypic analysis in field, such as precision robotics, production monitoring and crop productivity forecasts.*

Keywords: Deep Neural Networks; Tridimensional Reconstruction; Data Annotation; Classification and Segmentation; Point Clouds

1 INTRODUÇÃO

Este trabalho de Iniciação Científica se insere nas etapas finais do projeto *Agricultura ciente de ambiente: raciocínio sobre estrutura tridimensional no campo de cultivo (AACr3)*¹, cujo principal objetivo é utilizar os recentes avanços em visão computacional e algoritmos de aprendizagem profunda para automatizar o processo de recuperação e posterior detecção e classificação de objetos de interesse em cultivos, como plantas, folhas e frutos. Tais etapas correspondem às atividades: anotação em point clouds (nuvens de pontos), classificação e segmentação e avaliação da estimativa de características (como volume, biomassa e peso das frutas).

No início do projeto, estudou-se o trabalho anterior sobre o qual este está sendo desenvolvido: Santos et al. (2017). Imagens de videiras foram coletadas em campo por um *webcam* simples. A metodologia utilizada seguiu Mur-Artal et al. (2017) para a reconstrução de pose da câmera, *patch-based multiple-view stereo* (PMVS) para gerar nuvens de pontos que aproximassem a superfície externa dos objetos e um *support vector machine* (SVM) para classificação dos pontos tridimensionais em classes de interesse.

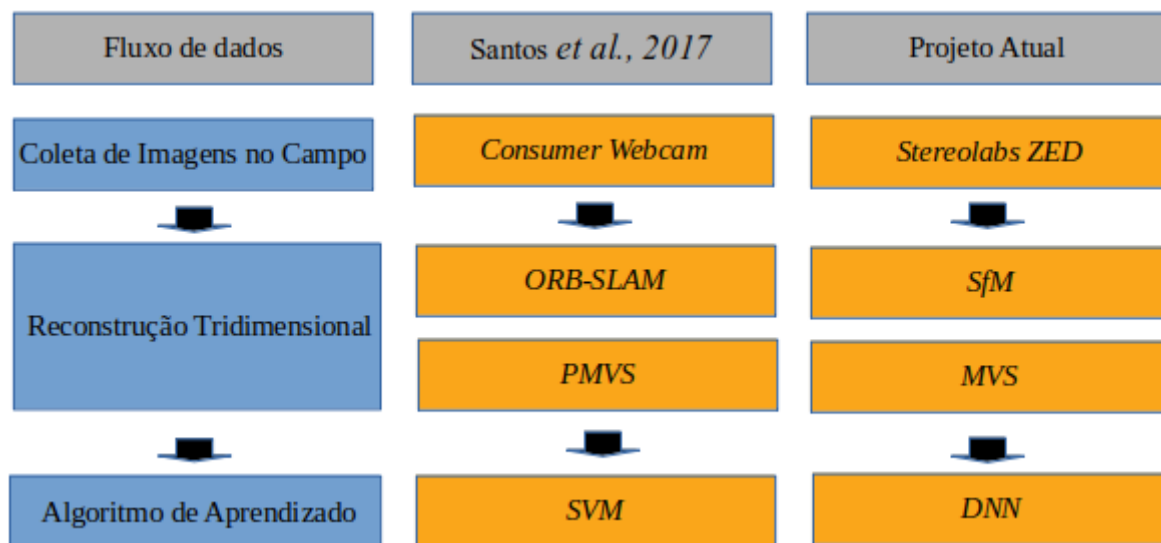


Figura 1. Comparação das etapas do fluxo de dados entre os projetos

Em nosso projeto, propusemos manter o fluxo de dados do trabalho anterior como exibido na Figura 1, alterando os algoritmos que geram as nuvens de pontos a partir de imagens e os que aprendem a classificá-las e segmentá-las.

Como alternativa ao SLAM e ao PMVS, utilizamos o COLMAP² (SCHÖNBERGER *et al.*, 2016) para gerar as nuvens de pontos. As reconstruções tridimensionais realizadas pelo programa podem ser divididas em dois momentos principais, que ocorrem em sequência: a reconstrução esparsa, resultado do algoritmo Structure from motion ou SfM, e a reconstrução densa, resultado do algoritmo *Multi View Stereo (MVS)*. Como referência teórica foi utilizado Hartley e Zisserman (2004).

O algoritmo de SfM utiliza conceitos da geometria epipolar para reconstruir a estrutura tridimensional de uma cena a partir de imagens. De modo simplificado, a reconstrução parte de correspondências de pontos entre duas imagens, ou seja, de projeções de um mesmo ponto do espaço tridimensional, estima uma matriz fundamental entre os planos de imagem, estima matrizes de câmera, as utiliza para reprojetar raios no espaço tridimensional e, finalmente, estima as coordenadas do ponto observado pelas duas imagens. Após um certo número de triangulações, o sistema realiza um refinamento conjunto dos pontos e poses de câmera (*Bundle Adjustment*) (TRIGGS *et al.*, 2000), buscando minimizar o erro de reprojeção. Os processos de extração de características e busca de correspondência ocorrem antes do SfM, em módulos separados do COLMAP. O algoritmo de MVS tem como objetivo gerar uma amostragem densa das superfícies a partir do resultado do SfM. Ele consiste numa etapa de *patch match stereo* para expandir as



correspondências e uma etapa que estima mapas de profundidade e normal das imagens, e funde-os numa reconstrução densa.

Como alternativa a SVM, propusemos o uso de redes neurais profundas, Deep Neural Network (DNN), algoritmos iterativos de aprendizado que buscam minimizar uma métrica de erro atualizando os parâmetros de suas camadas pelo algoritmo de backpropagation, como descrito em Chollet (2018). Detecção, classificação e segmentação são problemas importantes de Visão Computacional cujas soluções de estado da arte envolvem redes neurais profundas. Detecção é a tarefa de sinalizar se um dado objeto está em um conjunto de entrada. Classificação é a tarefa de escolher uma classe para o conjunto de entrada dentre um conjunto de classes. Segmentação é a tarefa de particionar a entrada em múltiplos segmentos de interesse. Essas partições também podem conter informação de classe e/ou instância (*instance segmentation*). Qi et al. (2016, 2017) detalham a arquitetura de uma rede neural profunda capaz de realizar classificação e segmentação em nuvens de pontos (PointNet).

O treinamento de redes neurais de aprendizado supervisionado depende da disponibilidade de dados anotados; no caso da PointNet, da disponibilidade de grande quantidade de nuvens de pontos anotadas em classes de objetos tridimensionais. A primeira pesquisa de ferramenta de anotação de nuvens de pontos foi o *SmartScenes ToolKit* (SSTK)³, utilizada em Dai et al. (2017). A ferramenta permitia a anotação de malhas tridimensionais em uma janela de um navegador de internet (como o Google Chrome), e provia uma ferramenta para transformar nuvens de ponto em malhas. Sua especificidade como ferramenta descentralizada, utilizada por diversos usuários pela internet, a necessidade de converter nuvens para malhas poligonais e grande curva de aprendizado indicaram que o desenvolvimento de uma aplicação de anotação própria do projeto era o mais adequado. Assim, foi iniciado o desenvolvimento da aplicação AACr3-Anotador, utilizando a biblioteca *Point Cloud Library* (PCL)⁴ e o framework Qt⁵.

A aplicação tem como objetivo permitir a visualização e manipulação de uma nuvem de pontos no espaço, permitindo a anotação manual de pontos pelo usuário em classes de interesse e/ou instâncias de classes. Como a quantidade de dados anotados necessários para treinar uma rede neural profunda é grande, seria inviável anotar as nuvens de modo totalmente manual. Propomos então que a ferramenta forneça uma pré-anotação de objetos da nuvem, a partir da

3 Disponível em: <<https://github.com/smartszenes/sstk>>.

4 Disponível em: <<https://pointclouds.org/>>.

5 Disponível em: <<https://www.qt.io/>>

projeção de segmentações bidimensionais nas imagens geradoras, geradas por uma rede neural de segmentação e classificação em imagens. Assim, não só seria poupado tempo do anotador humano, como também a reprojeção das anotações da nuvem (tanto manuais quanto geradas pela rede profunda) poderiam ser usados para treinar melhor a rede que atua na imagem.

2 METODOLOGIA

As imagens da vinícola foram coletados pela câmera ZED na Vinícola Gaspari, em Espírito Santo do Pinhal, SP. Elas foram capturadas em vídeos com resolução de 2208 por 1242 pixels, um para a lente esquerda e outro para a direita. A câmera também fornece a pose de sua lente esquerda (6DoF: rotação e translação, por odometria visual) e uma reconstrução do ambiente por meio de software proprietário, que se baseia somente nas imagens coletadas. De modo a acessar diretamente os dados armazenados no formato de saída da câmera (svo), foi utilizada a API Python do ZED SDK.



Figura 2 (a). Exemplo de imagem da lente esquerda da ZED (imagem à esquerda) **(b).** Exemplo de imagem da lente direita (imagem à direita)

A reconstrução da estrutura tridimensional de linhas da vinícola, representada em nuvens de pontos, a partir de imagens previamente coletadas em campo por uma câmera stereo *Stereolabs ZED*⁶ (Figuras 2a e 2b), foi gerada utilizando dois sistemas de reconstrução: o conjunto de software



da *ZED SDK*⁷, desenvolvido pelos fabricantes da câmera; e o *COLMAP* software *open-source* desenvolvido na ETH Zurich descrito nos dois artigos de Schönberger et al. (2016) Schönberger e Frahm (2016).

No intuito de explorar a geometria da câmera estéreo na reconstrução pelo COLMAP, pois o baseline (distância entre as lentes) facilita a triangulação, e como o software ainda não oferece suporte nativo a câmeras *stereo*, os parâmetros intrínsecos de cada câmera foram atualizados separadamente nas imagens, após a fase de extração de características.

Para comparar as reconstruções, as poses de câmera obtidas pela API do ZED SDK e pelo COLMAP foram registradas (alinhadas) pelo algoritmo ICP (ZHANG, 1994). O Iterative Closest Point (ICP) assume que duas nuvens de pontos têm a mesma forma e estão com orientação próxima, e tenta registrá-las iterativamente, minimizando o erro quadrático médio, a raiz da média da soma das distâncias entre vizinhos mais próximos. Selecionamos algumas comparações na seção Resultados.

A aplicação AACr3-Anotador utiliza as ferramentas de visualização e manipulação de nuvens de pontos da biblioteca PCL e interface gráfica construída pelo framework Qt, ambos controlados por lógica escrita em C++. Em seu estado atual de desenvolvimento, o programa realiza a projeção de pontos tridimensionais selecionados pelo usuário em pontos bidimensionais correspondentes nas imagens que foram utilizadas para gerar a nuvem. Tal funcionalidade foi implementada como teste para operações de projeção entre a nuvem e suas imagens geradoras. A operação de projeção segue a fórmula matricial abaixo, como descrito em Hartley e Zisserman (2004) :

$$x = PX \quad (1)$$

$$P = K[R \vee t] \quad (2)$$

A Equação 1 mostra que o ponto bidimensional na imagem (x) é produto da matriz de projeção (P) com o ponto tridimensional na nuvem. A Equação 2 detalha como é obtida a matriz de projeção por meio da matriz de parâmetros intrínsecos K, especificada pela ZED, e das matrizes de rotação R e translação t, obtidas na reconstrução pelo COLMAP.

3 RESULTADOS E DISCUSSÃO

3.1 Reconstruções

Até o momento, temos como resultados reconstruções de teste das linhas de uva feitas pelo ZED SDK e pelo COLMAP, e poses de câmeras das reconstruções. Eles são apresentados abaixo, por meio de uma tabela e imagens.

Foram feitas reconstruções com o ZED SDK para todos as linhas disponíveis, pois seu processamento dura um tempo curto. Elas apresentaram comportamentos variados: o modelo da primeira linha tem geometria e poses de câmera consistentes com as imagens fornecidas; o da segunda não reproduziu toda a largura da linha; o da terceira apresentou uma bifurcação da linha (Figura 3). As demais reconstruções encaixam-se em um dos casos descritos acima.

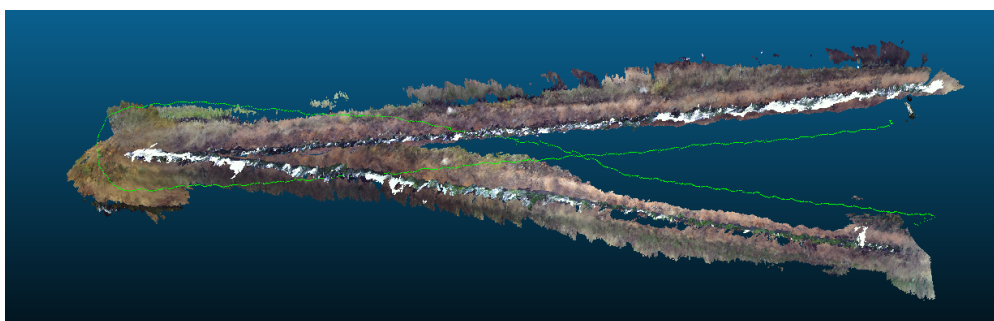


Figura 3. Cruzamento na trilha da terceira videira Cabernet Sauvignon, reconstruída com o ZED SDK: em verde, a trilha de câmera

As reconstruções do COLMAP são comparadas por meio da Tabela 1. Analisamos as escolhas de parâmetros buscando aqueles que promovam um número maior de pontos na nuvem. O primeiro parâmetro de interesse é o número de imagens fornecidas para a reconstrução: como foram capturadas em formato de vídeo, há muita redundância entre elas, dada a grande intersecção de conteúdo entre imagens adjacentes. Ele se destacou por ter a maior correlação com o número de pontos reconstruídos: as reconstruções esparsas com um sétimo das imagens, um terço das imagens e de todas as imagens capturadas têm 190.490, 483.222 e 663.899 pontos, respectivamente.

Há interesse em utilizar o menor número possível de imagens que garanta uma reconstrução com número de pontos suficiente para treinamento da rede neural, pois, para um grande de imagens de entrada, o Bundle Adjustment exibe complexidade de tempo linear no número de imagens registradas (SCHÖNBERGER et al., 2016). Em nossos experimentos, a reconstrução esparsa de um terço das imagens demorou aproximadamente um terço do tempo da reconstrução de todas as imagens.



Tabela 1. Comparação das reconstruções da linha um da uva Sauvignon realizadas com o COLMAP com diferentes parâmetros.

Tipo	# imagens	# pontos (Esparsa)	# pontos (Densa)	#matches
Terço	684	483.001	6.600.411	11.645
Sétimo	296	133.732	2.719.982	2.619
Terço Mesclado	694	483.222	10.605.697	23.660
Sétimo Mesclado	296	190.490	6.675.985	9.784
Terço Mesclado	694	485.173	---	23.310
Todas as Imagens Mescladas	2080	663.899	---	79.530
ZED SDK	0	0	330.551	----

O experimento de mesclar as imagens, ou seja, usar como entrada da reconstrução imagens da lente direita e esquerda alternadamente, teve grande efeito no número de correspondências (matches), como é visto comparando o exemplo do terço sem (11.645 matches) e com (23.660 matches) mesclagem. Usando um sétimo, a mesclagem aumentou consideravelmente o número de pontos, como pode ser visto na Tabela 1. Na Tabela 1, os tipos “Terço” e “Sétimo” se referem à fração do total de imagens utilizado. “Mesclado” se refere à técnica de mesclar as imagens das duas câmeras.

Com o COLMAP, a reconstrução esparsa mais bem-sucedida até o momento, usando como critério a relação entre o formato esperado da reconstrução e o tempo empregado no processamento, utilizou um terço das imagens, metade delas da lente esquerda e metade da direita (Figura 4). No entanto, ela ainda apresentou comportamentos inesperados: bifurcação na linha de uva e posição relativa inconsistente das câmeras esquerda e direita, ou seja, elas não são representadas como a estrutura rígida que são na realidade. Diferente do que ocorreu no caso do ZED SDK, a trilha da câmera não se cruzou e a bifurcação só ocorreu no final da trilha. A Figura 5 mostra o alinhamento de sua trilha com a do ZED SDK, que forneceu um RMS de 0.161897. As reconstruções densas da primeira e segunda linhas de videiras são apresentadas nas Figuras 6 e 7.

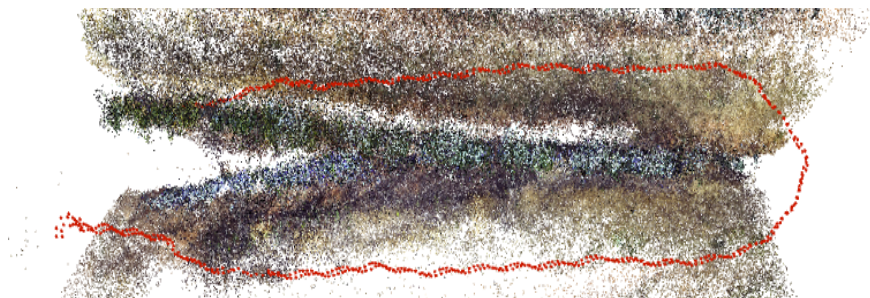


Figura 4. Reconstrução esparsa da primeira videira Cabernet Sauvignon com o COLMAP, utilizando um terço das imagens. É possível ver a bifurcação.

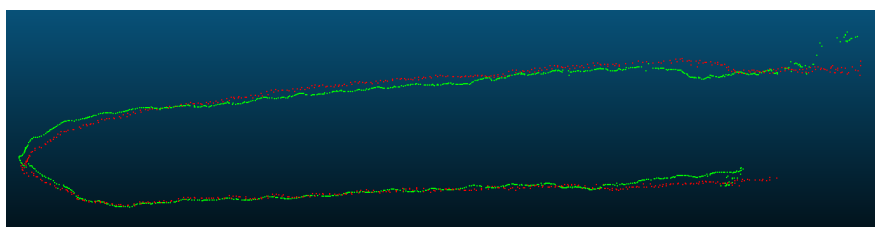


Figura 5. Comparação da trilha de câmera da reconstrução da Figura 3b (vermelho) com a trilha do ZED SDK (verde)

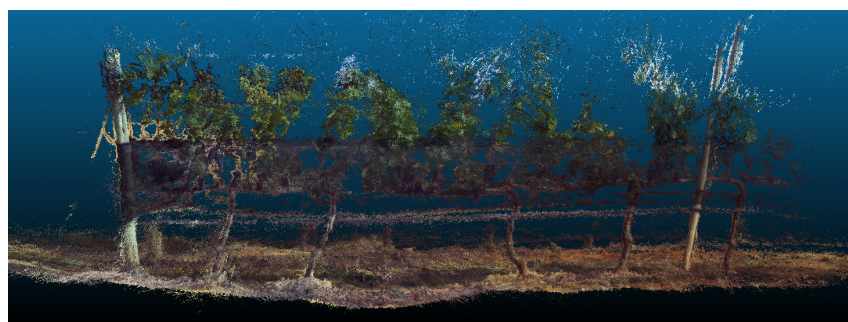


Figura 6. Reconstrução densa da primeira videira Cabernet Sauvignon com o COLMAP

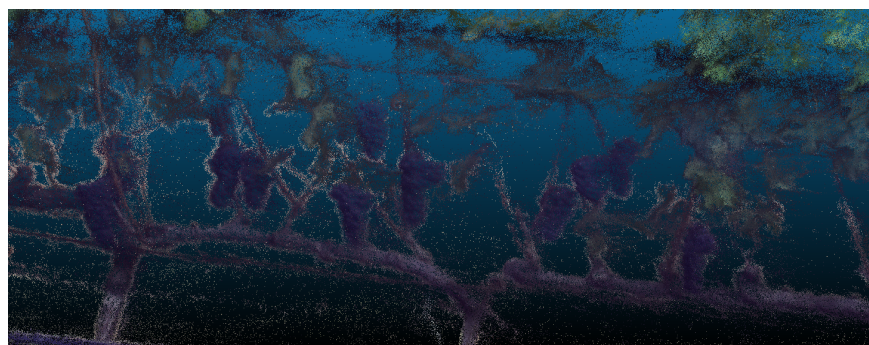


Figura 7. Detalhe da Reconstrução densa da segunda linha - note o detalhamento dos cachos de uva



O sucesso do aprendizado de uma rede profunda depende muito da oferta de grande quantidade de dados. As reconstruções finais obtidas possuem número de pontos da ordem de milhões (exemplo do “Terço Mesclado” com 10.605.697 pontos finais). Julgamos o volume de dados apropriado para o treinamento de uma rede como a PointNet, pois em seu artigo ela foi treinada e testada para segmentação semântica a partir da base de dados de Armeni et al. (2016), conjunto de 6.000m² de modelos de nuvens de pontos, extraíndo 4.096 pontos de cada metro quadrado para treino, utilizando total de 24.576.000 pontos para treino, escala comparável à de nossas nuvens.

Pelo critério quantitativo do número de pontos, e visualmente pelo detalhamento dos objetos de interesse, o COLMAP é a ferramenta mais adequada para se reconstruir as linhas de uva visando seu uso como entrada de algoritmos de aprendizado profundo quando comparada à ZED SDK, que produz nuvens muito esparsas.

Para a recuperação de poses de câmera, o ZED SDK é uma alternativa mais rápida em alguns casos, já que seus resultados não diferiram muito dos do COLMAP nos testes feitos até o momento. O ZED SDK usa odometria visual, uma técnica de reconstrução de poses de câmera incremental na qual a trilha é reconstruída imagem a imagem, o que pode causar drift devido ao erro acumulado a cada estimativa. Já o SfM utilizado pelo COLMAP otimiza a trilha de modo global, reduzindo o *drift* a cada iteração do Bundle Adjustment. Desse modo, a solução implementada pelo ZED SDK pode gerar poses absolutas muito diferentes do esperado quando comparadas às geradas pelo COLMAP e, portanto, não deve ser encarada como substituta à recuperação de trilha por SfM, mas sim uma aproximação.

Quanto à escolha de parâmetros para a reconstrução, só foi possível validar o número de imagens e mesclagem na melhora da qualidade e quantidade de pontos das reconstruções. O uso de parâmetros intrínsecos diferentes para as duas câmeras não produziu diferenças expressivas, provavelmente pela pequena diferença dos parâmetros.

3.2 Aplicação

O atual estado de desenvolvimento da aplicação permite a manipulação da nuvem de pontos e a seleção de um deles: o ponto selecionado é projetado na imagem geradora escolhida pelo usuário, caso ele seja visível por ela. A execução do programa ocorre em duas threads (linhas de execução), uma para o visualizador da nuvem e a outra para o controle do aplicativo. A Figura 8 explica a interface de usuário do programa.

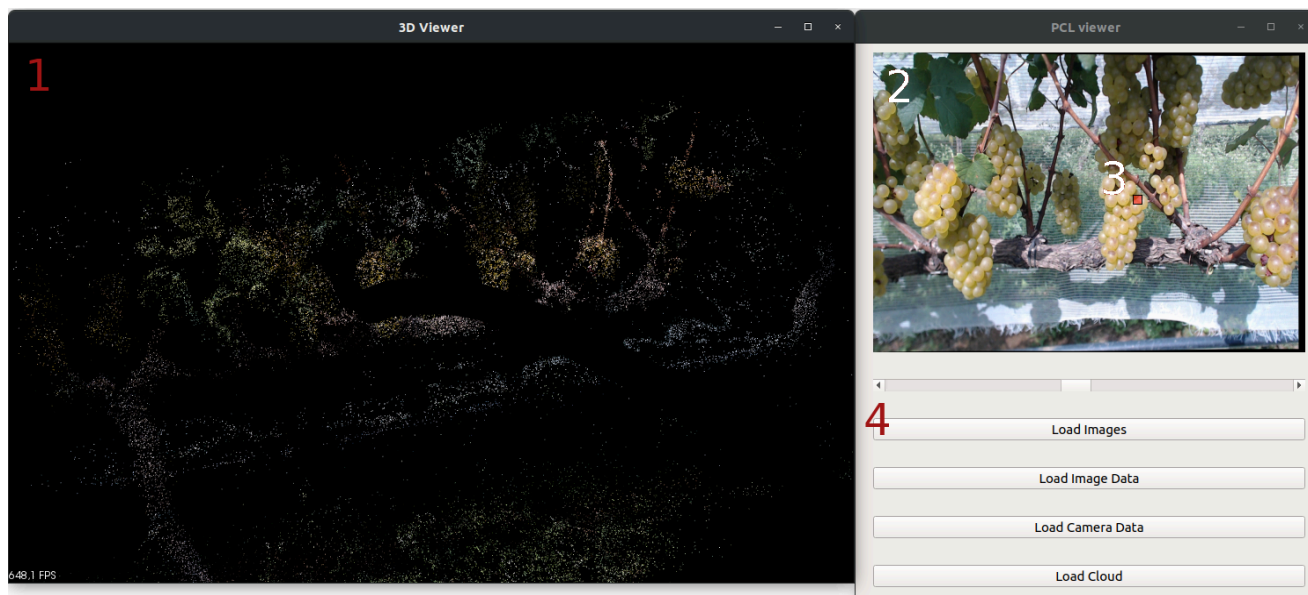


Figura 8. Interface de usuário da aplicação AACr3-Anotador. 1-Janela de manipulação da nuvem de pontos. 2-Janela de visualização das imagens. 3-Ponto selecionado na nuvem projetado na imagem. 4 – Botões para carregamento dos dados necessários para uso da ferramenta e escolha da imagem visualizada.

4 CONCLUSÕES

Após experimentos com as reconstruções, chegamos a parâmetros de reconstrução com o COLMAP que permitem processamento num tempo viável e produzem um número de pontos numa escala apropriada para treinar uma rede neural profunda. Reconstruções com a ZED SDK forneceram um modelo esparso demais para ser usado no treinamento de uma rede neural profunda, além de duvidarmos da acurácia de sua reconstrução de trilhas de câmera. Nas próximas etapas, a aquisição de um banco de imagens georreferenciado (com uso de GPS RTK com precisão centimétrica) irá permitir uma escolha melhor de parâmetros, baseada nas diferenças de trilhas de câmera, e a comparação de trilhas reconstruídas com um referencial real.

A função de projeção implementada na aplicação AACr3-Anotador apontou para a viabilidade do desenvolvimento de uma ferramenta própria de anotação. Na próxima etapa da Iniciação, será adicionada a função dual, ou seja, a reprojeção de pontos na imagem em pontos da nuvem tridimensional, permitindo assim a pré-anotação automatizada com uso de imagens previamente segmentadas e classificadas. Em seguida será implementada a interface de anotação manual das



nuvens. Por último será desenvolvida a interface que salva as nuvens anotadas em um formato adequado para o treinamento das redes neurais.

A última etapa da pesquisa se responsabilizará por treinar modelos de rede neural profundas para as tarefas de segmentação e classificação em nuvens de pontos.

5 AGRADECIMENTOS

Agradecemos à FAPESP pela concessão da Bolsa BCO – Iniciação Científica usufruída no período de 01/07/2019 a 30/06/2020, e sua renovação no período de 01/07/2020 a 28/02/2020.

6 REFERÊNCIAS

ARMENI, I. et al. **3D Semantic parsing of large-scale indoor spaces**. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 29., 2016, Las Vegas. **Proceedings...** Piscataway: IEEE, 2016. p. 1534-1543.

CHOLLET, F. **Deep learning with Python**. New York: Manning, 2018. 361 p.

DAI, A. et al. **ScanNet: richly-annotated 3D reconstructions of indoor scenes**. In: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 30., 2016, Honolulu. **Proceedings...** Piscataway: IEEE, 2017. p 2432-2443.

HARTLEY, R.; ZISSERMAN, A. **Multiple view geometry in computer vision**. 2. ed. Cambridge: Cambridge University Press, 2004.

MUR-ARTAL, R.; TARDÓS, J. D. **ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras**. IEEE Transactions on Robotics, New York, v. 33, n. 5, p. 1255-1262, Oct. 2017.

QI, C. R. et al. **Pointnet: deep learning on point sets for 3D classification and segmentation**. In: EUROPEAN CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 30., 2016, Honolulu. **Proceedings...** Piscataway: IEEE, 2017. p 77-85

QI, C. R. et al. **Pointnet++: deep hierarchical feature learning on point sets in a metric space**. In: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach. **Proceedings...** Red Hook: NY Curran Associates, [2018]. (Advances in neural information processing systems, 30).



SANTOS, T. T. et al. **Automatic grape bunch detection in vineyards based on affordable 3D phenotyping using a consumer webcam.** In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 11., 2017, Campinas. Ciência de dados na era da agricultura digital: anais. Campinas: Editora da Unicamp: Embrapa Informática Agropecuária, 2017. p. 89-98.

SCHÖNBERGER, J. L.; FRAHM, J. M. **Structure-from-Motion revisited.** In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 29., 2016, Las Vegas. Proceedings... Piscataway: IEEE, 2016. p. 4104-4113.

SCHÖNBERGER, J. L. et al. **Pixelwise view selection for unstructured multi-view stereo.** In: EUROPEAN CONFERENCE ON COMPUTER VISION, 14., 2016, Amsterdam. Proceedings... Cham: Springer: 2016. p. 501-518. (Lecture notes in computer science, 9907).

TRIGGS et al. **Bundle adjustment — a modern synthesis.** In: INTERNATIONAL WORKSHOP ON VISION ALGORITHMS, 1999, Corfu. Vision algorithms: theory and practice: proceedings. Berlin: Springer, 2000. p. 298–372. (Lecture notes in computer science, 1883)

ZHANG, Z. Y. **Iterative point matching for registration of free-form curves and surfaces.** International Journal of Computer Vision, Hingham, v. 13, n. 2, p. 119-152, Oct. 1994