12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

DETECÇÃO AUTOMÁTICA DE UVAS E FOLHAS EM VITICULTURA COM UMA REDE NEURAL YOLOv2

Andreza Aparecida dos **Santos**[1]**;** Sandra **Avila**[2]; Thiago Teixeira **Santos**[3]

Nº 18601

RESUMO *– Neste trabalho, o problema de detecção de frutas e folhas em viticultura para aplicações envolvendo sensoriamento próximo foi modelado como um problema de aprendizado supervisionado de máquina. Uma base de dados foi criada e manualmente anotada a partir de imagens obtidas em abril de 2017 na Vinícola Guaspari. No total são 11.883 imagens contendo exemplos de cachos de uvas e folhas. Uma rede convolutiva com arquitetura YOLOv2 foi treinada para localização e classificação de cachos e folhas. Testes quantitativos demonstraram resultados para a detecção e classificação com precisão de 100%, revocação de até 74,2% e F1-Score de 85,2% para classe "uva" e precisão de 100%, revocação de até 67,9% e F1-Score de 80,9% para a classe "folha". Testes qualitativos mostram que o modelo generaliza bem quando testado em fotografias de outras variedades de uvas. Esses resultados se mostram promissores para a melhoria do método e caminham para a possibilidade de aplicação em campo.*

Palavras-chaves: Detecção de frutos, Reconhecimento de Imagens, Viticultura, Aprendizagem profunda.

ABSTRACT – *In this work, we modeled the problem of detection of fruit and leaves in viticulture for proximal applications as a supervised machine learning task. We created and manually labeled a database of images obtained in April 2017 at Guaspari Winery. In total, the database consists of 11,883 images of bunch of grapes and leaves. We trained a convolutional network with YOLOv2 architecture to locate and classify bunch of grapes and leaves. Quantitative tests have shown results for detection and classification with precision of 100%, recall of 74.2% and F1-Score up to 85.2% for the class "grape" and precision of 100%, recall of 67.9% and F1-Score up to 80.9% for the class "leaf". Also, qualitative tests show that the model generalizes well when tested on photographs of other grape varieties. These results are promising and are moving towards the possibility of application in the field.*

Keywords: Fruit detection, Image Recognition, Viticulture, Deep Learning.

ISBN 978-85-7029-145-5

1. INTRODUÇÃO

A detecção e reconhecimento de frutos é um componente importante em aplicações de automação na área da agricultura de precisão, como técnicas de colheita automatizada, análise das plantas para correção de nutrientes e aplicação de insumos. Devido ao tamanho dos frutos, sensoriamento por satélite é inviável, de forma que o sensoriamento proximal se torna o mais apropriado. Nesse contexto, técnicas baseadas em imagens e redes neurais profundas se apresentam como o atual estado da arte para a tarefa de detecção e classificação de objetos (SA, I. et al., 2016).

Trabalhos na área da viticultura mostram a utilização de técnicas tradicionais de aprendizado de máquina a fim de detectar frutos e predizer safra (NUSKE et al., 2014) ou de detectar com precisão o momento de colheita de modo que seja possível a colheita através de robôs sem danificar a planta nem o fruto (LUO et al., 2016).

Nossos estudos iniciais utilizando técnicas de aprendizado de máquina para o problema de reconhecimento de bagas de uvas apresentaram até 79% de precisão (SANTOS; SANTOS, 2017a) com Máquinas de Vetores de Suporte combinadas com descritores de forma. A fim de melhorar os resultados, foram estudadas duas abordagens de aprendizado de máquina: por uma rede neural sequencial e uma rede neural convolutiva. Estas se mostraram eficazes no reconhecimento dos frutos, atingindo até 85% de precisão, revocação e F1-Score (SANTOS; SANTOS, 2017b), mas tais técnicas não realizavam a detecção espacial dos frutos, que é o foco do presente trabalho.

Como o objetivo final do projeto do qual esse estudo faz parte é realizar predições através de imagens vindas de uma câmera embarcada em um sistema móvel (VANTs, robôs, carros de serviço) seguindo as fileiras de videiras em campo, buscamos um modelo que fosse capaz de realizar predições em tempo real. Isso viabiliza aplicações que necessitam de atuação, em que a detecção e localização fazem parte do planejamento e tomada de decisão pelo sistema autônomo, o que também inviabiliza o uso das duas técnicas anteriores

Este artigo tem como objetivo investigar e analisar uma técnica de aprendizado supervisionado para detecção e classificação de imagens capaz de realizar predições em tempo real. Na seção de Material e Métodos apresentamos a base de dados, a arquitetura da rede YOLOv2 e as métricas utilizadas para avaliação dos resultados. Em Resultados e Discussão discutimos os resultados obtidos para os experimentos realizados. Na Conclusão apresentamos as contribuições do trabalho, as limitações do modelo e as possíveis linhas de pesquisa para trabalhos futuros.

ISBN 978-85-7029-145-5

2. MATERIAL E MÉTODOS

2.1. Aquisição da base de dados

As imagens utilizadas foram obtidas na Vinícola Guaspari (Espírito Santo do Pinhal – SP) em abril de 2017, por uma câmera digital SLR (Canon® EOS Rebel T3i) de lentes 18-55 milímetros configurada em modo automático. As imagens foram adquiridas de uma distância de aproximadamente 1 metro das plantas. As imagens apresentam frutos de uva da variedade Sirah em diferentes estágios de desenvolvimento e são coloridas, de 8 bits e possuem 1296 x 864 pixels, como ilustrado na Figura 1. No total são 59 fotografias. Apesar do número reduzido de fotografias, cada fotografia apresenta vários exemplos de cachos e folhas, permitindo a construção de um conjunto maior de amostras. Também foi mostrado em (SA, I. et al., 2016), que isso não se configura em um problema ao trabalharmos com redes neurais convolutivas. Essas redes apresentam propriedades como compartilhamento de parâmetros e equivalência de representações (GOODFELLOW, I. et al., 2016), permitindo aprendizado supervisionado com um número consideravelmente menor de amostras se comparadas a outras arquiteturas em redes neurais profundas.

Figura 1. Exemplo de imagens utilizadas para gerar a base de dados.

2.2. Anotação e Aumentação da Base de Dados

As amostras presentes na base de dados são compostas por *bounding boxes* contendo cachos de uvas (classe 0) ou folhas (classe 1). Essas *bounding boxes* foram criadas manualmente e anotadas a partir das fotografias originais. Assim, o número total de amostras é 842 (ver Figura 2).

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

Figura 2. Exemplos de *bounding boxes* anotadas do conjunto original de imagens.

Desse conjunto de fotografias, 20 foram separadas para compor o conjunto de teste, totalizando 223 amostras. O restante passou pelo processo de aumentação de dados. A base de dados foi aumentada utilizando a biblioteca imgaug (JUNG, A., 2018) em linguagem Python, de modo a gerar mais amostras diferentes (ver Figura 3). As técnicas de aumentação utilizadas foram: *scale, fliplr, flipud, dropout, gaussian blur, gaussian noise, add, rotate* e *crop and pad*. Essas técnicas foram aplicadas em conjunto e individualmente na parte das imagens da base da dados que serão utilizadas no conjunto de treinamento e validação do algoritmo. Ao final desse processo, o número de amostras acrescentadas na base de dados foi de 11.041, totalizando 11.883.

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

Figura 3. Exemplos de *bounding boxes* anotadas do conjunto aumentado de imagens.

2.3. Rede Neural Profunda YOLOv2

A rede objeto de estudo deste trabalho é chamada de "You Only Look Once" (REDMON, J. et al., 2017) na sua segunda versão (YOLOv2), ajustando-a para um problema de duas classes. Essa rede consegue operar a 45 frames por segundo, o que permite realizar predições em tempo real. A YOLOv2 realiza menos detecções erradas, reduzindo o número de falsos positivos, mas pode apresentar mais erros de localização. Comparada a outras redes que operam em tempo real (por exemplo, Mask R-CNN (HE, K. et al., 2017)), essa arquitetura e suas derivações se apresentam atualmente como as mais rápidas.

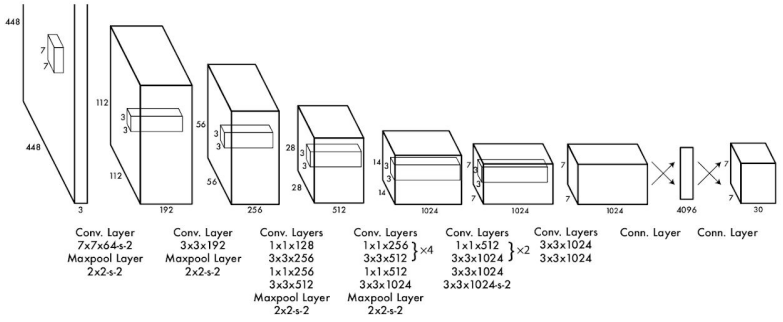


Figura 4. Arquitetura da rede. Imagem retirada de (REDMON, J. et al., 2016).

A arquitetura da rede YOLOv2 possui 24 camadas convolucionais seguidas de duas camadas *fully connected,* como mostra a Figura 4. Ela recebe como entrada uma imagem e a divide em uma grade de dimensão **S**x**S**. Caso o centro de algum objeto de interesse esteja dentro de uma das células dessa grade, essa célula fica responsável pela detecção do objeto. Cada célula da grade faz a predição de B *bounding boxes* e de um *índice de confiança*. O índice de confiança reflete a confiança do modelo de que existe um objeto dentro daquela *bounding box* e quão preciso ele a considera.

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

O índice de confiança é definido como *Prob(Object)* * *IoU*, onde *Prob(Object)* é a probabilidade de existir um objeto dentro da *bounding box* e *IoU* é a interseção sobre a união, calculado como mostrado na Figura 5. Assim, uma *bounding box* sem objeto deve apresentar índice de confiança igual a zero, caso contrário esse índice é igual a interseção sobre a união entre a *bounding box* presente no ground-truth e a predita. Como saída final a rede produz um tensor composto pela classe, índice de confiança e pelas coordenadas da *bounding box* predita. Para um problema de duas classes, a saída é um tensor de dimensões 7x7x12.

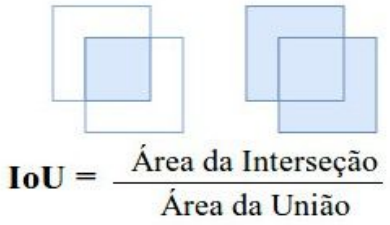


Figura 5. Ilustração referente ao cálculo da IoU.

O treinamento foi inicializado com os pesos de uma rede YOLOv2 pré-treinada por Redmon (REDMON et al, 2017) com a base ImageNet (RUSSAKOVSKY et al., 2015). Em seguida a rede foi treinada para nossa base de dados, utilizando *batches* com 32 imagens, ou seja, em cada passo do treinamento utilizamos 32 imagens diferentes do conjunto de treinamento, e a cada 10.000 iterações realizadas salvamos os pesos obtidos até o momento para posterior avaliação dos resultados. Durante o treinamento é possível observar o comportamento da média do erro obtido pela rede no conjunto de validação. Quando esse indicador começa a aumentar, há grandes chances de estar ocorrendo *overfitting* e o treinamento pode ser interrompido. Ao todo foram realizadas 90.000 iterações no treinamento.

A avaliação foi feita para o conjunto de pesos da iteração 90.000, que apresentou menor média de erro durante o treinamento. Utilizando esses pesos e as imagens contidas no conjunto de testes, nunca vistas pela rede antes, realizamos as predições da rede para esse conjunto de imagens e salvamos em arquivo as saídas da rede. O processo de aquisição de resultados foi realizado 5 vezes alterando-se o valor do limiar dentro do intervalo [0,5, 0,9]. Esse limiar determina se o resultado a predição entra ou não no conjunto final de respostas, caso o índice de confiança seja maior ou igual ao limiar, a predição é considerada no resultado final da rede.

Passamos então para a etapa de cálculo das quantidades de verdadeiro positivo, falso negativo e falso positivo. Um verdadeiro positivo é observado quando uma *bounding box* detectada pela rede tem probabilidade da classe certa acima do limiar *t* e possui uma *IoU* de 0,5. Caso contrário, ela é marcada como um falso positivo. As *bounding boxes* do conjunto e anotação que não foram associadas a nenhum verdadeiro positivo é marcada como falso negativo. Também foram calculados os índices de precisão (*P*), revocação (*R*) e F1-Score para cada classe, seguindo as fórmulas mostradas na Equação 1.

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

$$P = \frac{Vp}{Vp+Fp}, \quad R = \frac{Vp}{Vp+Fn}, \quad F1-Score = \frac{2 \cdot P \cdot R}{P+R} \qquad (1)$$

onde *Vp* é o número de verdadeiro positivo (detecções corretas), *Fp* é o número de falsos positivos (detecções erradas) e *Fn* é o número de falsos negativos (falha de detecção).

3. RESULTADO E DISCUSSÃO

As tabelas abaixo mostram os resultados dos testes realizados para cada valor de limiar testado. A Tabela 1 é referente aos resultados para a classe "uva" e a Tabela 2 é referente aos resultados para a classe "folha".

Tabela 1. Quantidade de verdadeiros positivos, falsos negativos, falsos positivos, precisão, revocação e F1-Score para a classe "uva" para valores de limiar de 0,5 a 0,9.

Limiar	Verdadeiros Positivos	Falsos Negativos	Falsos Positivos	Precisão	Revocação	F1-Score
0,5	95	33	0	1	74,22%	85,20%
0,6	93	34	0	1	73,23%	84,55%
0,7	83	44	0	1	65,35%	79,05%
0,8	60	67	0	1	47,24%	64,17%
0,9	1	126	0	1	0,79%	1,56%

Tabela 2. Quantidade de verdadeiros positivos, falsos negativos, falsos positivos, precisão, revocação e F1-Score para a classe "folha" para valores de limiar de 0,5 a 0,9.

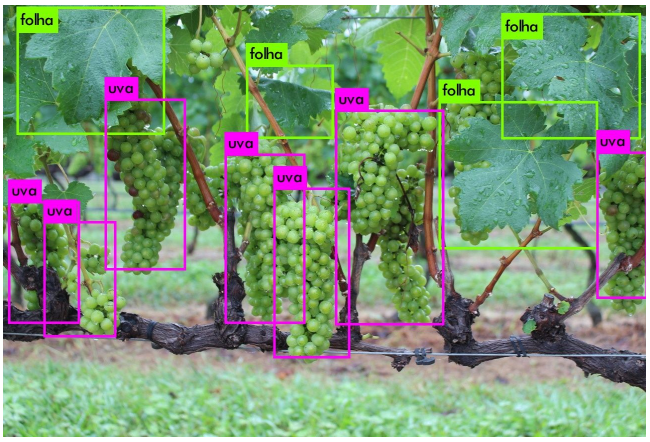
Limiar	Verdadeiros Positivos	Falsos Negativos	Falsos Positivos	Precisão	Revocação	F1-Score
0,5	72	34	0	1	67,92%	80,9%
0,6	68	38	0	1	64,15%	78,16%
0,7	57	49	0	1	53,77%	69,94%
0,8	33	73	0	1	31,13%	47,48%
0,9	4	102	0	1	3,77%	7,27%

Podemos observar que, para o limiar de 0,5 o número total de predições foi de 234, que é maior que o número total de amostras no conjunto de teste. Também é possível verificar que essa diferença ocorre devido a uma predição a mais feita para uma amostra da classe "uva". O valor correto de 233 amostras se estabiliza a partir do limiar 0,6, pois com o aumento do limiar falsos positivos são filtrados, indicando que elas apresentavam um baixo índice de confiança da classe e foram filtradas. Do limiar 0,5 para o 0,6 ocorre uma queda no índice do F1-Score da classe "uva" para 84,55% devido à dupla predição.

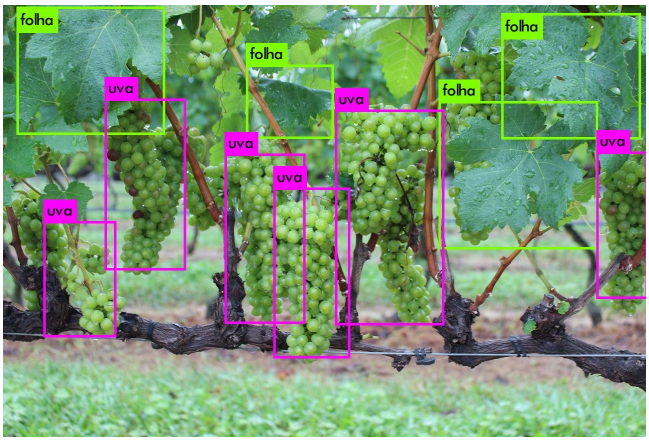
ISBN 978-85-7029-145-5

Observando a partir do limiar 0,6 temos um aumento crescente no número de falsos negativos, o que indica que para algumas amostras presentes na base de dados, a rede não conseguiu fazer a predição com índice de confiança acima do limiar aceitável. Isso gera uma queda no índice F1-Score para 1,56% para o limiar 0,9 na classe “uva”. Dessa forma, temos os melhores resultados apresentados para um limiar de 0,6, que não possui nenhuma amostra erroneamente marcada como positiva.

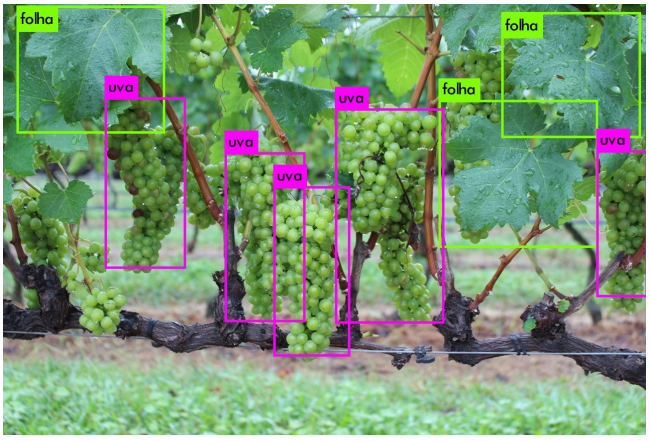
É possível atribuir os resultados obtidos à falta de diversidade nas amostras, que apesar da grande quantidade, foram obtidas a partir de um conjunto de fotografias bem comportadas. Aumentando-se a base de dados com exemplos que abranjam uma variação maior de defeitos, de modo a treinar a rede expondo-a às imperfeições do mundo real, pode vir gerar resultados melhores.



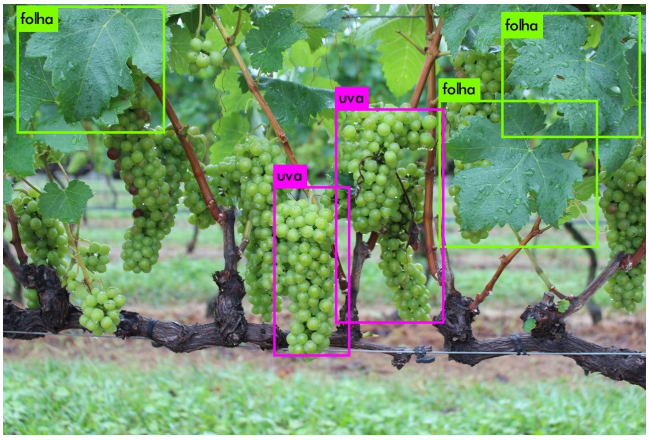
(a)



(b)



(c)



(d)

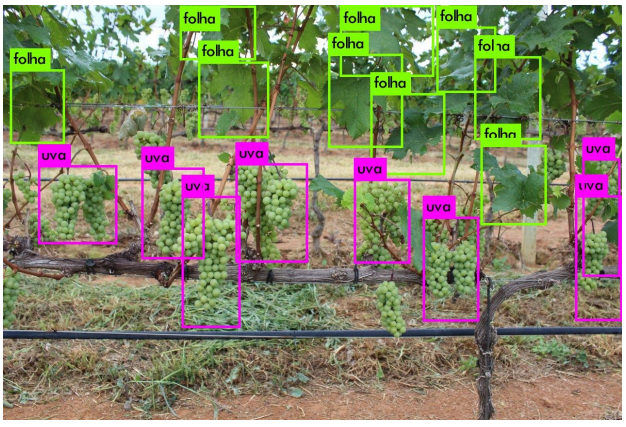
12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5

(e)

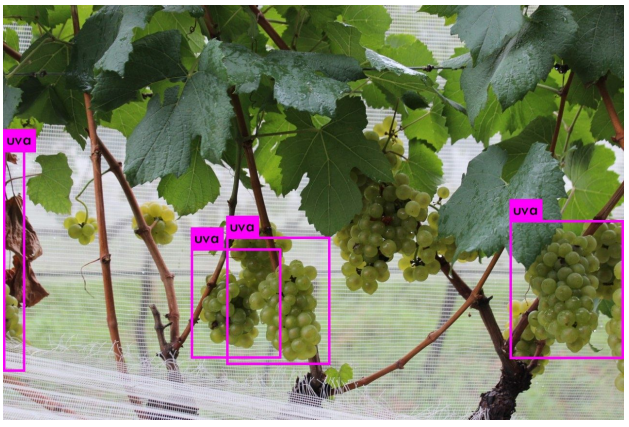
Figura 6. Resultados das predições variando o limiar para uma mesma imagem. Os limiares estão em ordem crescente de (a) 0,5 a (e) 0,9.

A Figura 6 mostra alguns resultados da predição da mesma imagem para diferentes valores de limiares. Podemos ver que ao aumentarmos os valores do limiar, a rede passa a detectar menos objetos, produzindo uma imagem com menor quantidade de *bounding boxes* detectadas, mostrando que os índices de confiança das probabilidades preditas ainda podem ser melhorados.

Como forma de avaliar se o modelo consegue generalizar, isso quer dizer, se consegue aplicar o aprendizado adquirido com o treinamento em amostras de uvas da variedade Syrah em outras variedades de uvas nunca vistas antes, realizamos alguns testes qualitativos como mostra a Figura 7.

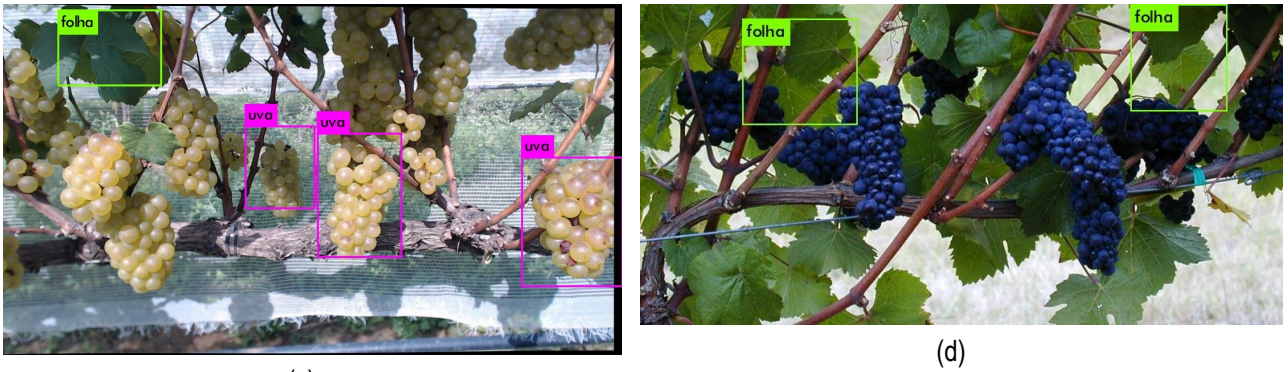


(a)



(b)

12º Congresso Interinstitucional de Iniciação Científica – CIIC 2018
01 a 03 de agosto de 2018 – Campinas, São Paulo
ISBN 978-85-7029-145-5



(c) (d)

Figura 7. Resultados das predições para amostras de diferentes variedades de uvas. (a) Cabernet, (b) Chardonnay e (c) Chardonnay e (d) Pinot Noir. (a),(b),(c) fotografias tiradas na Vinícola Guaspari (d) fotografia de autoria de David Stutz retirada da internet.

Podemos ver que, para variedades que possuem uma coloração parecida com a da variedade na qual a rede foi treinada como em (a), (b) e (c), a YOLOv2 produz resultados muito próximos àqueles obtidos para as amostras do conjunto de teste. Já em (d), como a variedade Pinot Noir possui uma coloração muito diferente e nunca vista pela rede anteriormente, a YOLOv2 não é capaz de detectar esses cachos. Vale notar que as fotografias (b), (c) foram obtidas utilizando uma câmera Logitech C920 e a (d) foi obtida utilizando uma Olympus C2000Z, duas câmeras diferentes da câmera utilizada na criação do conjunto que compõe a base de dados. Conseguir realizar boas predições em fotografias geradas por câmeras distintas, portanto com qualidades de imagem diferentes, é bom um indicativo da capacidade de generalização do modelo.

4. CONCLUSÃO

Nesse trabalho apresentamos o problema de detecção e classificação de uvas e folhas e o abordamos a partir de uma técnica de aprendizado supervisionado utilizando a rede neural convolutiva YOLOv2. No geral a rede neural YOLOv2 apresentou resultados promissores para o problema de detecção apresentado. Devido a sua característica de realizar predições a partir de *bounding boxes*, existe uma limitação no número de elemento próximos distintos que ela consegue detectar, o que pode se apresentar como uma limitação devido à disposição física da planta objeto do estudo.

Como trabalho futuro, pretendemos investigar formas de melhorar o desempenho dessa rede a fim de colocá-la para operar em campo. Uma vez que a YOLOv2 apresentou resultados promissores, outro caminho é avaliar a “Fast YOLO” apresentada em (REDMON, J. et al., 2016), que é uma versão menos profunda da YOLOv2, mas que opera a 155 frames por segundo devido a menor quantidade de camadas convolucionais, e a rede chamada “YOLOv3” (REDMON, J. et al., 2018), que é a terceira versão da YOLO com modificações para melhorar as predições. Outra rede

ISBN 978-85-7029-145-5

interessante de se investigar é a “Mask R-CNN” (HE, K. et al., 2017) que é capaz de produzir máscaras bem ajustadas aos objetos preditos.

5. AGRADECIMENTOS

A autora agradece ao PIBIC/CNPq pela bolsa concedida (#161165/2017-6), à Embrapa pela oportunidade de desenvolver o projeto, à Vinícola Guaspari por nos receber e apoiar, ao Programa de Concessão de GPUs da NVIDIA (*NVIDIA’s GPU Grant Program*) por ceder uma GPU Titan X que foi essencial para a realização deste estudo, à Microsoft por ceder o uso de GPUs no Azure, e aos seus orientadores Thiago e Sandra pelos conhecimentos e ensinamentos compartilhados e também pelas risadas durante o processo.

6. REFERÊNCIAS

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.. Deep learning. **The MIT Press**, Cambridge, MA, 2016.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R.. Mask r-cnn. In: **Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE**, 2017. p. 2980-2988.

JUNG, A. Imgaug. Disponível em: <http://imgaug.readthedocs.io/en/latest/>. Acessado em: 12 jun. 2018.

LUO, L.; TANG, Y.; ZOU, X.; YE, M.; FENG, W.; LI, G.. Vision-based extraction of spatial information in grape clusters for harvesting robots. **Biosystems Engineering**, v. 151, p. 90-104, 2016.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You Only Look Once: Unified, real-time object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 779-788.

REDMON, J.; FARHADI, A. YOLO9000: better, faster, stronger. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,** 2017. p. 6517-6525.

REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. **arXiv preprint arXiv:1804.02767,** 2018.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; DENG, A.. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211-252, 2015.

SA, I.; GE, Z.; DAYOUB, F.; UPCROFT, B.; PEREZ, T.; MCCOOL, C. DeepFruits: A Fruit Detection System Using Deep Neural Networks. **Sensors**, vol. 16, n. 8, p. 1222. 2016.

ISBN 978-85-7029-145-5

SANTOS, A. A. dos; SANTOS, T. T.(a) Detecção de frutos em campo por aprendizado de máquina. In: Congresso Interinstitucional de Iniciação Científica, 11., 2017, Campinas. **Anais**. [S.l: s.n.], 2017. p. 1-9. Disponível em: Disponível em: <https://www.bdpa.cnptia.embrapa.br/consulta/busca?b=ad&id=1077535>. Acessado em: 13 jun. 2018.

SANTOS, A. A. dos; SANTOS, T. T.(b) Estudo de métodos de aprendizagem profunda para reconhecimento de bagas de uva**.** In: Mostra de Estagiários e Bolsistas da Embrapa Informática Agropecuária, 13., 2017, Campinas. **Resumos expandidos**. Brasília, DF: Embrapa, 2017.. p. 43-46. Disponível em: <https://www.bdpa.cnptia.embrapa.br/consulta/busca?b=ad&id=1085142>. Acessado em: 13 jun. 2018.

NUSKE, S.; WILSHUSEN, K.; ACHAR, S.; YODER, L.; NARASIMHAN, S.; SINGH, S. Automated visual yield estimation in vineyards. **Journal of Field Robotics**, v. 31, n. 5, p. 837-860, 2014.