Um Sistema de Reconhecimento de Espécies de Moscas-das-Frutas

Matheus Macedo Leonardo[1], **Fabio Augusto Faria**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal do São Paulo (Unifesp)
São José dos Campos – SP – Brasil

matheus.macedo.leonardo@gmail.com,ffaria@unifesp.br

Resumo. *Moscas-das-frutas são insetos de grande importância biológica e econômica para a agricultura de diferentes países no mundo. As perdas diretas e indiretas causadas por essa praga podem exceder USD 2 Bilhões, tornando-a um dos grandes problemas para a agricultura mundial, especialmente para o Brasil. Nesse contexto, o desenvolvimento de sistemas para a identificação automática ou semiautomática de moscas-das-frutas do gênero Anastrepha pode auxiliar os especialistas (entomólogos) na redução de tempo de análise e nas perdas. Neste trabalho de iniciação científica, nós propusemos um sistema de reconhecimento de moscas-das-frutas baseada em duas diferentes representações de imagens: (1) representação de nível-médio (Bag-of-Words); (2) representação de aprendizagem profunda (deep learning-based features). Ambos trabalhos conseguiram excelentes resultados de eficácia para a tarefa de identificação de três espécies das moscas-das-frutas do gênero Anastrepha superando os resultados encontrados na literatura.*

Introdução

As moscas-das-frutas pertencem a família *Tephritidae*, que abrange aproximadamente 5,000 espécies. Estão distribuídas por todo o mundo e várias espécies são importantes pragas agrícolas. Os danos são causados pelas larvas que se alimentam no interior da fruta, e a tornando imprópria para consumo e comercialização. Além disso, algumas espécies também são de importância quarentenária, assim essas espécies dificultam a comercialização internacional de frutas frescas (in natura). Os países em que as pragas quarentenárias não ocorrem impõem barreiras alfandegárias para a importação de mercadorias do país onde a praga está presente.

Entre as moscas-das-frutas de importância econômica nas Américas estão as espécies do gênero *Anastrepha*. Esse gênero é o mais diversificado nos trópicos e subtrópicos da América com aproximadamente 300 espécies conhecidas, na qual 120 são registradas no Brasil [Zucchi, R. A. 2008]. No entanto, poucas espécies são de importância econômica no Brasil. Chamadas de moscas-das-frutas da América do Sul *Anastrepha fraterculus* (Wiedemann), moscas-das-frutas da Índia Ocidental *Anastrepha obliqua* (Macquart),e moscas-das-frutas da goiaba *Anastrepha striata* Schiner, são três espécies consideradas pragas de importância quarentenária pelas agências reguladoras.

A identificação das espécies é uma etapa crucial para o desenvolvimento de estudos na biologia. A identificação das espécies de *Anastrepha* são baseadas em padrões das asas, e principalmente nos acúleos (a parte perfurante do ovipositor da fêmea). No entanto, as diferenças entre as espécies de alguns complexos de moscas-das-frutas são

difíceis de delimitar. *Anastrepha fraterculus* é o caso mais emblemático de um complexo de espécies crípticas nas Américas, por ser uma praga importante somente em algumas áreas de sua ocorrência, na qual abrange do México ao norte da Argentina [Schutze et al. 2017]. Portanto, erros na identificação podem ser um problema sério para a implementação de restrições de quarentena, gerenciamento integrado de pragas e outros programas de controle [McPheron 2000].

Novos processos de identificação de insetos estão sendo usados como análises morfométricas e moleculares para uma identificação precisa das moscas-das-frutas do gênero *Anastrepha* [Bomfim et al. 2011, Bomfim et al. 2014]. No entanto, Martineau et al. [Martineau et al. 2017] apontou que poucos trabalhos na literatura propuseram identificar o gênero *Anastrepha* através de processamento de imagem e técnicas de aprendizagem de máquina. Provavelmente esse fato está relacionado a alta semelhança entre as espécies pertencentes ao gênero *Anastrepha*.

Neste sentido, este trabalho de iniciação propôs a utilização de uma abordagem de representação de nível-médio (*bag of visual words* - BossaNova [Avila et al. 2013]) baseada em descritores locais de imagem e representação de aprendizagem profunda baseada em arquiteturas de redes neurais convolucionais (CNN) para a tarefa de identificação de moscas-das-frutas do gênero *Anastrepha* [Leonardo et al. 2017, Leonardo et al. 2018]. Além disso, foi comparada a efetividade dos resultados de diferentes técnicas de aprendizagem de máquina para auxiliar o desenvolvimento de um sistema de tempo real para identificação de espécies do gênero *Anastrepha*. Esse sistema mostrou ser uma boa solução para a tarefa alvo.

Sistema de Reconhecimento de Moscas-das-frutas

Esta seção aborda a explicação de cada um dos dois métodos propostos neste trabalho de iniciação científica.

Representação de Nível-médio (*Bag-of-Words*)

A classificação de imagens é tipicamente abordada em três etapas, (I) extração de características visuais locais, que consiste no processo de extração de informações diretamente dos pixeis da imagem, (II) extração de características de nível-médio, faz com que as informações sejam generalizadas, agregando abstração ao modelo, e por fim (III) classificação supervisionada, consiste em uma técnica de aprendizagem de máquina que permite a extração de um modelo geral a partir dos dados. Esta seção mostra as etapas do sistema de reconhecimento de espécies de moscas-das-frutas baseado em representação de nível-médio.

A Figura 1 mostra o sistema de reconhecimento de espécies de moscas-das-frutas baseado em representação de nível-médio. Para isso, a abordagem BossaNova [Avila et al. 2013] foi adotada.

O sistema inicia com a extração local de características, onde o descritor local SIFT [Lowe 2004] realiza a extração através dos pontos de interesses detectados pelo algoritmo FAST [Rosten and Drummond 2005]. Então, uma extração de características de nível-médio é realizada por meio do uso a abordagem BossaNova, em que vetores de características SIFT dos pontos das imagens da base são agrupados para originar o dicionário visual de palavras (dimensões do histograma). Uma etapa de treino do modelo

de decisão é realizada em que os vetores BossaNova das imagens do conjunto de treino são utilizados como entrada para treinar uma técnica de aprendizagem de máquina. Finalmente, a etapa de previsão do modelo de decisão em que o modelo treinado utiliza os vetores do BossaNova das imagens do conjunto de teste para prever suas classes. Mais detalhes sobre a abordagem BossaNova, podem ser encontrados em [Avila et al. 2013].

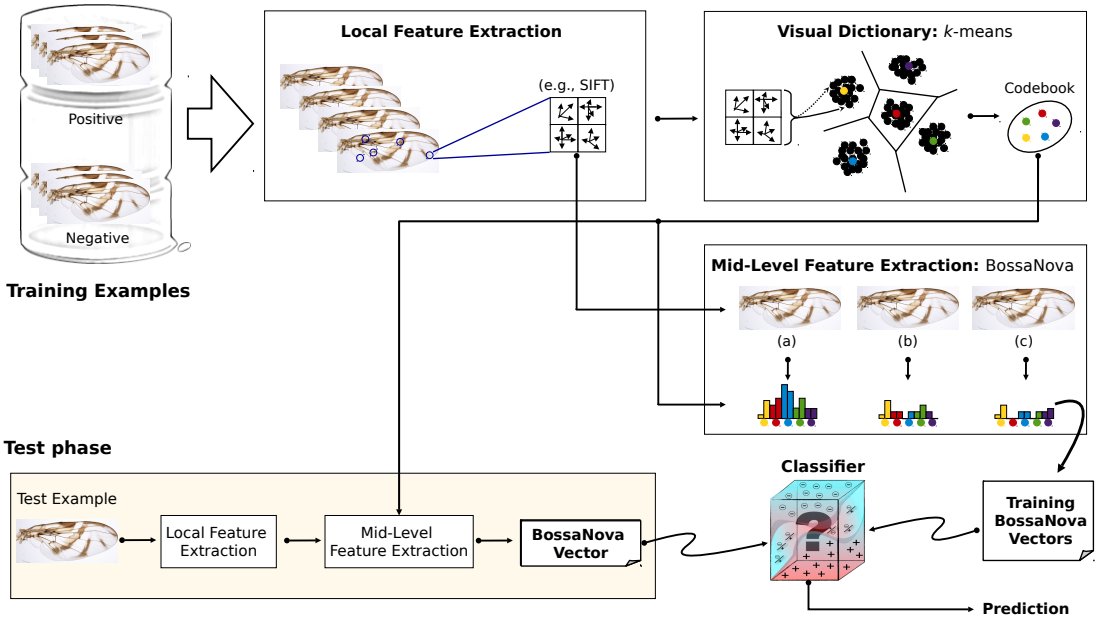

Figura 1. O *pipeline* principal da abordagem BossaNova [Leonardo et al. 2017].

Representação de Aprendizagem Produnda (*Deep Learning*)

Nesta seção serão apresentadas algumas das mais bem sucedidas famílias de arquiteturas de aprendizagem profunda utilizadas na literatura (VGG, GoogLeNet e ResNet) baseadas em camadas convolucionais [Leonardo et al. 2018].

Visual Geometry Group (VGG)

As arquiteturas VGG [Simonyan and Zisserman 2014] com 16 camadas (VGG16) and 19 camadas (VGG19) foram desenvolvidas no *Visual Geometry Group* (VGG) para a competição ImageNet de 2014. O time VGG obteve o primeiro e segundo lugar na tarefa de localização e classificação, respectivamente. Essas arquiteturas são estruturada inicialmente com cinco blocos de camadas convolucionais seguidas por três camadas completamente conectadas. As camadas convolucionais usam 3×3 *kernels* com *stride* de 1 e *padding* de 1 para assegurar que cada mapa de ativação contenha as mesmas dimensões da camada anterior. Uma função de ativação ReLU (*rectified linear unit*) está após cada camada convolucional e uma operação de *max pooling* é usada no final de cada bloco para redução da dimensionalidade espacial. As camadas de *max pooling* usam *kernels* 2×2 com um *stride* de 2 e sem *padding* para garantir que cada dimensão espacial do mapa de ativação da camada anterior seja reduzida pela metade. Duas camadas completamente conectadas com 4096 valores calculados pela função de ativação ReLU são gerados antes da

camada final *softmax* de 1000 nós/classes. A desvantagem entre as arquiteturas VGG16 e VGG19 é tempo de treinamento, uso de memória e escolha de parâmetros. VGG16 tem aproximadamente 138 milhões de parâmetros e a VGG19 tem 143 milhões de parâmetros.

GoogLeNet

A arquitetura GoogLeNet foi apresentada como GoogLeNet (Inception V1), depois surgiu a Inception V2 e atualmente existe a Inception V3 [Szegedy et al. 2016]. Os módulos Inception são extratores de características convolucionais responsáveis por aprender representações ricas com poucos parâmetros. Uma camada convolucional tradicional tenta aprender filtros em espaço 3D por meio de 2 dimensões espaciais (altura e largura) e 1 dimensão de canal. Assim, um único *kernel* está encarregado de mapear simultaneamente correlações entre canis e correlações espaciais. A ideia por trás do módulo Inception é tornar esse processo mais fácil e eficiente ao fatorar explicitamente uma série de operações que examinam independentemente as correlações entre canais e as correlações espaciais. Já a arquitetura Xception [Chollet 2016] é uma extensão da arquitetura Inception com a substituição do módulo Inception padrão por convolução separável *depthwise*. Ao invés de particionar os dados de entrada em várias pedaços comprimidos, o novo módulo mapeia as correlações espaciais para cada canal de saída separadamente e então, aplica 1×1 convolução separável *depthwise* para capturar a correlação entre canais. Assim, esse processo procura primeiro por correlações 2D, seguido por correlações 1D, resultando em um mapeamento completo 3D (2D+1D). A arquitetura Xception supera ligeiramente a arquitetura InceptionV3 na base ImageNet e supera amplamente o desempenho em base maior com 17.000 classes, com número similar de parâmetros e maior eficiência. O Xception tem 22.855.952 parâmetros treináveis, enquanto o Inception V3 tem 23.626.728 parâmetros treináveis.

ResNet

Residual Networks (ResNets) [He et al. 2016] são redes convolucionais profundas, as quais têm como ideia básica pular blocos de camadas convolucionais usando conexões de atalho para formar blocos chamados blocos residuais. Esses blocos residuais empilhados melhoram muito a eficiência do treinamento e resolvem o problema de degradação das características extraídas ao longo da rede convolucional. Na arquitetura ResNet-50, os blocos básicos seguem duas simples regras: (i) para mesmo tamanho de mapa de características de saída, as camadas têm o mesmo número de filtros; and (ii) se o tamanho do mapa de características é reduzido pela metade, o número de filtros é dobrado. A amostragem ou *down-sampling* é realizada diretamente pelas camadas convolucionais que têm um *stride* de 2 e uma normalização de *batch* é realizada logo após cada convolução e antes da função de ativação ReLU. O número total de camadas é 50 com 23534.592 parâmetros treináveis.

Formas de Treinamento de Arquiteturas de Aprendizagem Profunda

Na literatura, poucos trabalhos realizam treinamento completo (em inglês, *learning from scratch*) de redes neurais convolucionais (em inglês, Convolutional Neural Networks – CNN). Este fato ocorre pela insuficiente quantidade de dados da aplicação alvo que permita um ajuste desejável do modelo de aprendizagem. A alternativa adotada pelos pesquisadores é a estratégia de transferência de aprendizagem (em inglês, *transfer learning* [Yosinski et al. 2014]), a qual se utiliza de arquiteturas CNN pré-treinadas em uma grande base de dados (e.g. ImageNet) para classificar dados de uma aplicação específica ou utiliza dessas arquiteturas CNN como um poderoso descritor de imagem da aplicação alvo. Dentre as formas de realizar transferência de aprendizagem estão:

- **CNN como descritor de imagem:** Utiliza-se uma CNN pré-treinada com a base ImageNet, remove-se as últimas camadas completamente conectada e assim, utiliza-se a arquitetura CNN como um extrator de características de imagens para nova aplicação. Cada arquitetura CNN codifica as propriedades visuais das imagens para um vetor de característica com tamanho específico (e.g., VGG16/VGG19 com 4096 dimensões, Inception com 1024 dimensões e Resnet-50 com 2048 dimensões). Uma vez extraídas as características das imagens da aplicação alvo, utiliza-se esses vetores de características como entrada para um método de aprendizagem de máquina (e.g. SVM Linear [Boser et al. 1992]).
- **Ajuste fino da CNN (em inglês, fine-tuning):** Esta estratégia não é apenas substituir ou retreinar o classificador na última camada da CNN para a nova aplicação, mas também refinar os pesos de algumas camadas de uma arquitetura CNN pré-treinada na ImageNet pela estratégia de retro propagação de erro (em inglês, *back-propagation*). Essa etapa de refinar os pesos da arquitetura pode ser realizada em todas as camadas ou apenas em algumas delas, mantendo fixos os pesos das primeiras camadas e ajustando pesos das camadas restantes. A motivação para refinar pesos apenas de parte das camadas da CNN se dá pelo fato que as camadas iniciais extraem propriedades visuais mais genéricas das imagens (e.g. bordas, cantos, junções e cores) que devem ser úteis para qualquer aplicação. Já as últimas camadas da arquitetura CNN buscam propriedades visuais mais específicas da base de dados da aplicação alvo [Larochelle et al. 2009].
- **Modelos pré-treinados:** Desde que arquiteturas CNN podem levar muito tempo para treinar utilizando-se de múltiplas GPUs na base ImageNet, algo comum que ocorre na literatura é encontrar arquiteturas já treinadas e com os pesos já refinados para o uso. Essa pode ser uma outra estratégia de aprendizagem disponível que basta o usuário substituir a última camada completamente conectada que é específica para ImageNet (1000 classes) por uma camada com o número de nós igual ao de classes da nova aplicação.

Bases de Dados

Em nossos experimentos, nós utilizamos imagens microscópica de três espécimes de *Anastrepha* da coleção do Instituto Biológico de São Paulo. Este base, chamada de base de imagem original (O), é composta de 301 imagens (resolução 2560×1920) e dividido em três categorias diferentes, *A. fraterculus* (100), *A. obliqua* (101) e *A. sororcula* (100) [Leonardo et al. 2017].

A. fraterculus *A. obliqua* *A. sororcula*

Figura 2. Exemplos de asas das espécies estudas [Faria et al. 2014].

Tabela 1. Resultados de eficácia (em %) entre todos os classificadores baseados em nível-médio. [Leonardo et al. 2017]

Descritor	Técnicas de Aprendizagem de Máquina								
	MLP	NB	DT	NBT	kNN1	kNN3	kNN5	SL	SVM
BRIEF	**92.0**	**73.5**	**79.8**	78.8	86.4	**90.0**	**88.1**	82.7	90.7
BRISK	**87.4**	51.5	74.4	69.1	78.4	74.4	71.4	79.7	87.0
FREAK	**88.4**	55.8	67.8	70.4	76.7	75.1	73.7	**84.0**	85.0
ORB	90.0	61.8	75.1	74.4	86.7	84.4	82.4	82.1	**90.4**
SIFT	**84.7**	53.8	62.5	61.8	67.2	68.4	68.5	75.1	84.4
SURF	**89.4**	63.5	62.5	70.1	77.4	78.7	79.7	66.8	87.4
F-SIFT	**94.7**	62.1	76.7	**82.1**	**87.4**	85.1	84.4	82.4	**93.7**
F-SURF	**84.7**	49.2	65.8	66.1	76.1	74.1	72.1	80.4	83.4
Média	**88.9**	58.9	70.6	71.6	79.5	78.8	77.5	79.2	**87.7**
IC	**2.4**	5.5	4.7	4.57	4.8	5.0	4.9	4.0	**2.5**

Resultados e Discussões

Nesta seção serão mostradas análises divididas em quatro partes. Primeiro, experimentos realizados no artigo científico publicado na conferência e-Science [Leonardo et al. 2017] para verificarmos o comportamento do sistema de reconhecimento baseado em representação de nível-médio (Seção 2.1). Segundo, os experimentos contidos no artigo científico publicado na conferência SIBGRAPI [Leonardo et al. 2018] são mostrados os resultados com representação de aprendizagem profunda. Finalmente, realizamos uma comparação entre os melhores classificadores utilizando as representações nível-médio, de aprendizagem profunda e os métodos encontrados na literatura.

Classificadores baseados em Representação Nível-médio

Nesta seção, nós analisamos oito descritores locais baseados em pontos de interesses (BRIEF [Calonder et al. 2010], BRISK [Leutenegger et al. 2011], FREAK [Alahi et al. 2012], ORB [Rublee et al. 2011], F-SIFT, F-SURF, SIFT [Lowe 2004], and SURF [Bay et al. 2008]); e nove técnicas de aprendizagem de máquina (MLP, NB, DT, NBT, kNN1, kNN3, kNN5, SL, and SVM).

A Tabela 1 mostra os resultados de eficácia para todos os classificadores baseados em nível-médio para o protocolo *5-fold cross validation*. Em azul estão os melhores descritores de imagem para cada técnica de aprendizagem de máquina. As células cinzas estão as melhores técnicas de aprendizagem de máquina para cada descritor de imagem.

Primeiramente, nós podemos observar que os descritores BRIEF e F-SIFT conseguiram quatro dos melhores resultados de eficácia entre nove técnicas de aprendizagem de máquina (em azul). O descritor FREAK conseguiu um melhor resultado usando a técnica *simple logistic* (SL). Além disso, nós podemos notar que o descritor BRIEF conseguiu a melhor acurácia média ($84,7\%$).

Ainda, como pode ser observado a técnica *multilayer perceptron* (MLP) conseguiu sete melhores resultados entre os oitos descritores locais (célula cinza) usados neste trabalho. A técnica SVM conseguiu um melhor resultado com 90,4% de acurácia média usando descritor ORB. A técnica MLP usando descritor F-SIFT foi a melhor tupla (descritor+técnica de aprendizagem) com $94,7\%$ de acurácia média (em azul e célula cinza). Finalmente, nós podemos verificar que as técnicas MLP e SVM foram as melhores com acurácia média de $88,9\%$ e $87,7\%$, respectivamente.

Classificadores baseados em Representação Profunda

A Tabela 2 mostra os resultados de eficácia entre cinco arquiteturas *deep learning* (Inception, ResNet, VGG16, VGG19, and Xception) e nove técnicas de aprendizagem de máquina (DT, kNN1, kNN3, kNN5, kNN7, MLP, NB, SGD, SVM) para um protocolo *5-fold cross-validation*. Em azul estão as melhores representações *deep features* para cada técnica de aprendizagem de máquina. As células cinzas estão as melhores técnicas de aprendizagem de máquina para cada representação *deep learning*.

Tabela 2. Resultados de eficácia (em %) entre os classificadores baseados em representação de aprendizagem profunda [Leonardo et al. 2018].

Deep Learning	**Técnicas de Aprendizagem de Máquina**								
	DT	**KNN1**	**KNN3**	**KNN5**	**KNN7**	**MLP**	**NB**	**SGD**	**SVM**
Inception	57,83	71,09	70,09	70,09	66,77	87,7	54,13	88,03	**88,37**
ResNet	65,43	75,08	77,08	77,74	80,06	89,03	73,40	89,70	**90,36**
VGG16	**75,75**	84,39	**89,02**	**87,71**	**87,73**	**95,02**	**75,75**	**93,36**	**95,68**
VGG19	72,1	**84,72**	82,08	80,09	81,39	92,68	67,13	91,35	**94,34**
Xception	51,48	60,79	56,12	55,77	55,44	68,33	51,82	78,41	**78,74**

Nós podemos observar que as representações extraídas da arquitetura VGG16 conseguiram oito dos melhores resultados entre as nove técnicas de aprendizagem de máquina disponíveis (em azul). Além disso, nós podemos notar que a técnica SVM usando *kernel* linear conseguiu os melhores resultados em todas as cinco representações profundas (célula cinza) utilizadas neste trabalho. Finalmente, podemos comentar que a tupla SVM+VGG16 foi a melhor neste trabalho com 95,68% de acurácia média (em azul e célula cinza).

Comparação entre os Melhores Classificadores

Nesta seção nós comparamos os melhores classificadores baseados em nível-médio (F-SIFT+MLP [Leonardo et al. 2017]), classificadores baseados em representação de aprendizagem profunda (Inception+SVM, ResNet+SVM, VGG16+SVM, VGG19+SVM, and Xception+SVM) e o método estado-da-arte (LCH+SVM [Faria et al. 2014]. LCH+SVM é uma técnica SVM com *kernel* polinomial usando descritor de imagem chamado *Local Color Histogram* [Swain and Ballard 1991]. F-SIFT+MLP é uma técnica MLP usando representação em nível-médio [Avila et al. 2011] com detector de pontos de interesse FAST [Rosten and Drummond 2006] e descritor local SIFT [Lowe 2004].

A Figura 3 mostra os resultados de eficácia entre as melhores tuplas (representação + técnica de aprendizagem de máquina) e o melhor método existente na literatura. Apesar do VGG16+SVM (em azul) ter conseguido a melhor acurácia média (95,68%), quando calculamos o intervalo de confiança com nível de significância de 0.05, é possível observar que não existe diferença estatística entre os classificadores. Entretanto, é importante notar que a abordagem LCH+SVM conseguiu bons resultados de eficácia extraindo propriedades de cor de imagens melhoradas (e.g., segmentação e operações morfológicas). Portanto, esta abordagem baseada em cor não pode ser utilizada em sistemas reais, diferente das outras técnicas (F-SIFT+MLP and VGG16+SVM).

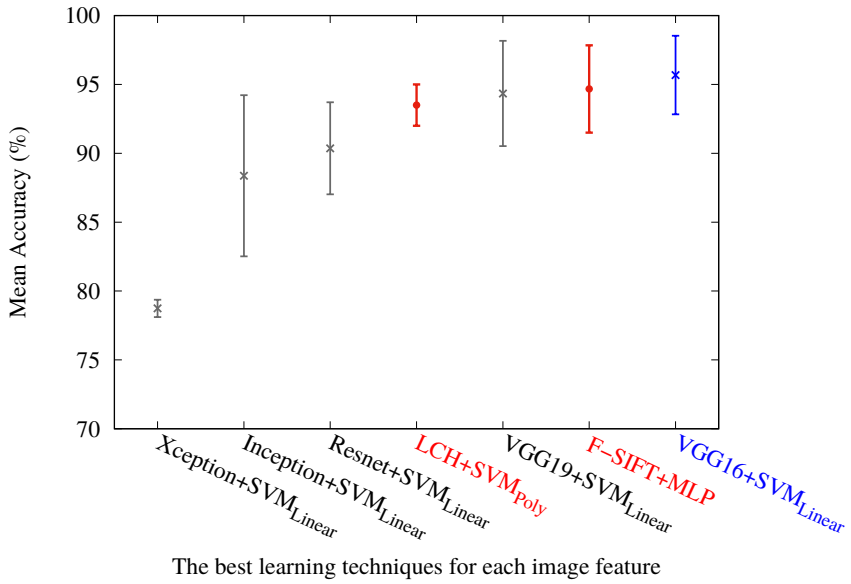

Figura 3. Resultados de eficácia para cada característica de imagem com intervalo de confiança de 95% (IC), i.e, um nível de significância de 0.05. Em azul está SVM$_{Linear}$ usando características VGG16 que conseguiu a melhor acurácia média [Leonardo et al. 2018].

Conclusão

Neste trabalho de iniciação científica, nós propomos um sistema de reconhecimento de moscas-das-frutas com duas diferentes representações (nível-médio e aprendizagem profunda). Diferente técnicas de aprendizagem de máquina, descritores de imagens e arquiteturas de aprendizagem profunda foram comparadas conseguindo um excelente resultados de 95,68% de acurácia média utilizando a abordagem VGG16+SVM na aplicação alvo. Esta abordagem consegue o melhor resultado sem qualquer operação adicional de melhoramento de imagem. Este fato é muito importante considerar na construção de um sistema de tempo real que poderá ajudar os escassos especialistas (entomólogos) na luta contra essas pragas de plantação. Como trabalhos futuro, nós pretendemos realizar experimentos com outras espécies de moscas-das-frutas e técnicas de aprendizagem de máquina. Outro trabalho pode ser o desenvolvimento de um sistema móvel para auxiliar os especialistas nos seus trabalhos de campo.

Agradecimentos

Os autores agradecem o CNPq pelo apoio financeiro por meio do Projeto Universal (408919/2016-7), o prof. Dr. Roberto Zucchi (Esalq/USP), profa. Dra. Sandra Ávila (IC/Unicamp), prof. Dr. Tiago Carvalho (IFSP-Campinas) e Msc. Edmar Rezende pela colaboração.

Referências

Alahi, A., Ortiz, R., and Vandergheynst, P. (2012). FREAK: Fast retina keypoint. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517.

Avila, S., Thome, N., Cord, M., Valle, E., and de A. Araújo, A. (2011). BOSSA: Extended BoW formalism for image classification. In *IEEE International Conference on Image Processing*, pages 2909–2912.

Avila, S., Thome, N., Cord, M., Valle, E., and de A. Araújo, A. (2013). Pooling in image representation: the visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359.

Bomfim, Z., Lima, K., Silva, J., Costa, M., and Zucchi, R. (2011). A morphometric and molecular study of Anastrepha pickeli Lima (Diptera: Tephritidae). *Neotropical Entomology*, 40:587–594.

Bomfim, Z., Lima, K., Silva, J., Costa, M., and Zucchi, R. (2014). Morphometric and Molecular Characterization of Anastrepha Species in the spatulata Group (Diptera, Tephritidae). *Annals of the Entomological Society of America*, 5:893–901.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, pages 144–152.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*, pages 778–792.

Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*.

Faria, F., Perre, P., Zucchi, R., Jorge, L., Lewinsohn, T., Rocha, A., and da S. Torres, R. (2014). Automatic identification of fruit flies (diptera: Tephritidae). *Journal of Visual Communication and Image Representation*, 25(7):1516–1527.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.*, 10:1–40.

Leonardo, M. M., Avila, S., Zucchi, R. A., and Faria, F. A. (2017). Mid-level image representation for fruit fly identification (diptera: Tephritidae). In *2017 IEEE 13th International Conference on e-Science (e-Science)*, pages 202–209.

Leonardo, M. M., Carvalho, T. J., Rezende, E., Zucchi, R., and Faria, F. A. (2018). Deep feature-based classifiers for fruit fly identification (diptera: Tephritidae). In *31st Conference on Graphics, Patterns and Images (SIBGRAPI 2018)*.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision*, pages 2548–2555.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., and Venturini, G. (2017). A survey on image-based insect classification. *Pattern Recognition*, 65(C):273–284.

McPheron, B. A. (2000). Population genetics and cryptic species. *Area-wide Control of Fruit Flies and Other Insect Pests*, pages 483–490.

Rosten, E. and Drummond, T. (2005). Fusing points and lines for high performance tracking. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1508–1515.

Rosten, E. and Drummond, T. (2006). Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, pages 430–443.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision*, pages 2564–2571.

Schutze, M. K., Virgilio, M., Norrbom, A., and Clarke, A. R. (2017). Tephritid integrative taxonomy: Where we are now, with a focus on the resolution of three tropical fruit fly species complexes. *Annual Review of Entomology*, 62(1):147–164.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zucchi, R. A. (2008). Fruit flies in Brazil: *Anastrepha* species and their host plants and parasitoids. http://www.lea.esalq.usp.br/anastrepha/.